

La memoria come finzione

Una teoria formale della ricostruzione mnestica

Proemio: La finzione come struttura necessaria

La parola "finzione" porta con sé un'ambiguità che conviene dissolvere subito, prima che avveleni l'intero percorso. Nel senso comune, e nella letteratura psicologica che ha dominato il campo dai lavori di Elizabeth Loftus in poi, la finzione mnestica è un *difetto*: la memoria produce racconti falsi perché è un sistema difettoso, soggetto a interferenze, suggestionabile, incapace di registrare fedelmente l'esperienza. Il programma di questa monografia è opposto: sostenere che la memoria produce finzione non *nonostante* la sua struttura ma *a causa* di essa, che la finzione è la forma necessaria che la memoria assume quando ottimizza l'obiettivo per cui è costruita — un obiettivo che non è l'accuratezza della registrazione ma la coerenza generativa dell'identità.

Questa tesi non è nichilistica. Non implica che la memoria sia inaffidabile in modo uniforme, né che non esista distinzione tra ricordi corretti e falsi. Implica qualcosa di più preciso: che la distinzione tra ricordo fedele e ricordo fabbricato non è la distinzione primaria da fare, e che un sistema che ottimizzasse la fedeltà a discapito della coerenza generativa non sarebbe una memoria migliore — sarebbe qualcosa di strutturalmente diverso, con funzioni cognitive radicalmente diverse.

Il modello che si propone è il seguente. La memoria non è un archivio che conserva rappresentazioni di eventi passati per recuperarle al bisogno. È un *sistema inferenziale generativo* che mantiene un modello compresso del passato — nel senso tecnico del capitolo precedente: un programma p con sofisticazione alta e residuo d ridotto — e che, quando interpellato, *ricostruisce* l'esperienza passata campionando dalla distribuzione implicita nel modello. Il ricordo è sempre, in questa accezione, un'istanza generata dal modello, non una fotocopia dell'originale. La generazione introduce necessariamente varianza: ogni richiamo produce un'istanza leggermente diversa, influenzata dalle condizioni di campionamento — lo stato corrente dell'agente, il contesto in cui il richiamo avviene, le credenze disponibili al momento della generazione.

La finzione mnestica è dunque la varianza del processo generativo — non l'errore di un archivio difettoso, ma la firma computazionale di un sistema che genera piuttosto che recuperare. Questa distinzione ha conseguenze formali precise, che costituiscono il programma dei capitoli che seguono.

Il percorso si articola in sei movimenti. Il primo costruisce il modello formale della memoria come compressore lossy con struttura generativa, e ne deriva le proprietà matematiche elementari. Il secondo affronta il riconsolidamento — la riscrittura del programma generativo ogni volta che un ricordo viene richiamato — come processo di ottimizzazione dinamica con una funzione obiettivo che non è la fedeltà ma la coerenza predittiva. Il terzo analizza la narrativa come codec di compressione: il modo in cui la struttura narrativa abbassa KC dell'esperienza autobiografica al costo di una distorsione sistematica. Il quarto formalizza la confabulazione come inferenza bayesiana sotto una distribuzione a priori mal calibrata, e ne ricava una previsione sulla struttura degli errori che è verificabile. Il quinto estende l'analisi alla memoria collettiva e alle sue dinamiche di consolidamento e revisione. Il sesto ritorna, con gli strumenti acquisiti, al confronto tra memoria umana e sistemi artificiali — un confronto che la monografia precedente ha aperto e che qui si approfondisce da una prospettiva diversa.

Capitolo I — Il modello formale: memoria come compressore lossy generativo

I.1 Perché un archivio non basta

Il modello intuitivo della memoria come archivio — un luogo in cui i ricordi vengono depositati e da cui vengono recuperati — è falsificato da un insieme di evidenze sperimentali che la psicologia cognitiva ha accumulato con coerenza crescente negli ultimi ottant'anni. Bartlett, negli anni Trenta, mostrò che il ricordo di storie è ricostruttivo: i soggetti non recuperano la storia come fu letta, ma la *ricostruiscono* a partire da uno schema — uno schema culturalmente determinato che riempie le lacune, elimina i dettagli incoerenti, produce una versione più coerente dell'originale. Neisser, cinquant'anni dopo, documentò che il ricordo autobiografico di eventi pubblici — l'attentato a Kennedy, il disastro dello Challenger — è sistematicamente distorto nella direzione di ciò che si sapeva *prima* dell'evento, non di ciò che si apprese *durante*. Loftus ha mostrato che basta una domanda suggestiva per impiantare ricordi di eventi che non sono mai accaduti.

Tutte queste evidenze convergono su un fatto: la memoria non recupera, genera. Ma la letteratura si è fermata qui — sul fatto empirico — senza costruire un modello formale di cosa significhi "generare" in questo contesto, quali siano le proprietà matematiche del processo generativo, e perché un tale sistema evolutivamente ottimizzato debba generare piuttosto che conservare.

I.2 La memoria come distribuzione condizionale

Si propone il seguente modello formale. Sia Ω lo spazio degli eventi esperienziali — le situazioni percettive, emotive e cognitive che un agente può attraversare. Un evento e

$\in \Omega$ ha una rappresentazione computazionale — una stringa x_e che lo codifica con sufficiente fedeltà per gli scopi cognitivi dell'agente.

La memoria non conserva x_e . Conserva un modello compresso m_e — nel senso della decomposizione modello-residuo del Capitolo VII della monografia precedente — che cattura la struttura non-accidentale di x_e e lascia perdere il residuo d_e . Formalmente:

$$m_e = p_{\{x_e\}} \quad (\text{il modello ottimale di } x_e \text{ nel senso della sofisticazione})$$

Quando l'agente viene interpellato — quando si produce una *cue* di recupero c — la memoria non restituisce m_e direttamente. Campiona una stringa $\hat{x}$ dalla distribuzione condizionale:

$$P(\hat{x} \mid m_e, c, \theta)$$

dove θ rappresenta lo stato corrente dell'agente — le sue credenze, il suo stato emotivo, il contesto in cui il recupero avviene. Il ricordo $\hat{x}$ è un'istanza campionata da questa distribuzione: è *compatibile* con il modello m_e (nel senso che m_e può generare $\hat{x}$ tramite U), ma non è deterministicamente determinata da m_e . La stessa *cue* c in stati θ diversi produce istanze $\hat{x}$ diverse.

La varianza del ricordo — la misura di quanto le diverse istanze $\hat{x}$ prodotte dallo stesso modello m_e in condizioni diverse si discostino l'una dall'altra — è la *varianza di generazione*, e corrisponde formalmente alla varianza del campionamento dalla distribuzione $P(\cdot \mid m_e, c, \theta)$.

I.3 La funzione obiettivo: non la fedeltà ma la coerenza predittiva

Perché il sistema dovrebbe ottimizzare una distribuzione generativa invece di conservare x_e direttamente? La risposta sta nella funzione obiettivo del sistema.

Un archivio ottimizza la *fedeltà di recupero*: minimizza la distanza $d(\hat{x}, x_e)$ tra il ricordo e l'evento originale. Un sistema generativo ottimizza qualcosa di diverso: la *coerenza predittiva* del modello m_e con l'esperienza futura dell'agente. Formalmente, si definisce la funzione obiettivo come:

$$J(m_e) = -\mathbb{E}_{\{c, \theta\}} [\log P_{\text{agent}}(\text{esito} \mid \hat{x}, \text{azione})]$$

dove P_{agent} è la distribuzione di probabilità dell'agente sugli esiti delle proprie azioni, e $\hat{x}$ è il ricordo generato da m_e dato c e θ . J misura quanto il modello m_e contribuisce a calibrare correttamente le aspettative dell'agente — quanto i ricordi generati da m_e portano a decisioni buone.

Un modello m_e che conservi fedelmente x_e non ottimizza necessariamente J : se x_e contiene molto rumore accidentale — dettagli irrilevanti per le decisioni future — conservarli non migliora le previsioni. Peggio: conservare il residuo d_e può *peggiore* J se il residuo è specifico al contesto originale e non si generalizza a contesti futuri. Il sistema generativo elimina d_e e ottimizza m_e verso la struttura di x_e che è predittivamente rilevante — verso ciò che dell'evento passato informa sulle strutture del mondo che si ripetono.

Questo è il fondamento evolutivo della memoria come compressore lossy: un sistema che conserva la struttura predittivamente rilevante e lascia perdere il residuo accidentale è, in ambienti strutturati, più adattivo di un sistema che conserva tutto. La finzione — la varianza di generazione, la deviazione dal ricordo "fedele" — è il prezzo pagato per la generalizzabilità del modello.

I.4 La connessione con il principio di energia libera

Il modello generativo della memoria converge, da una direzione diversa, con il *principio di energia libera* proposto da Karl Friston come teoria unificata della cognizione biologica. Il principio afferma che i sistemi cognitivi biologici minimizzano la *sorpresa* — tecnicamente, l'energia libera variazionale F , che è un upper bound sulla log-probabilità negativa delle percezioni sotto il modello generativo interno:

$$\begin{aligned} F &= \mathbb{E}_q [\log q(s) - \log p(o, s \mid m)] \\ &= KL[q(s) \parallel p(s \mid o, m)] + \log p(o \mid m) \end{aligned}$$

dove $q(s)$ è la distribuzione approssimata sugli stati nascosti s , $p(o, s \mid m)$ è il modello generativo con parametri m , e KL è la divergenza di Kullback-Leibler. Minimizzare F equivale simultaneamente a massimizzare l'evidenza del modello (il termine $\log p(o \mid m)$) e a minimizzare la divergenza tra la distribuzione approssimata e la distribuzione a posteriori vera.

La connessione con il modello mnestico è la seguente. Il modello generativo m nella formula di Friston corrisponde, nel caso mnestico, al modello compresso m_e . Il processo di campionamento del ricordo $\hat{x}$ corrisponde alla inferenza variazionale sullo stato nascosto s . La funzione obiettivo J del paragrafo precedente è una versione semplificata di $-F$: massimizzare la coerenza predittiva è massimizzare l'evidenza del modello generativo.

La divergenza KL nella formula di Friston cattura il costo dell'approssimazione: la differenza tra ciò che il modello prevede e ciò che l'agente può effettivamente computare. Nel caso mnestico, questa divergenza corrisponde alla distorsione introdotta dal processo di campionamento — la distanza tra il ricordo generato $\hat{x}$ e la distribuzione "vera" degli eventi compatibili con m_e . Minimizzare questa divergenza è minimizzare la varianza di generazione, ma sotto il vincolo computazionale che la distribuzione approssimata q sia trattabile. Il compromesso tra fedeltà e trattabilità è formalmente identico al compromesso tra KC e Kt della monografia precedente: la

memoria ottimizza la profondità predittiva compatibilmente con le risorse computazionali disponibili.

I.5 La lossy compression come inevitabilità formale

Si dimostra ora che la lossy compression — la perdita di informazione rispetto all'evento originale — è un'inevitabilità formale e non un difetto contingente del sistema.

Si supponga per assurdo che la memoria conservi x_e senza perdita: $m_e = x_e$. Allora $KC(m_e) = KC(x_e)$, e la profondità $D_s(m_e) = D_s(x_e)$. Per un evento esperienziale tipico, x_e include una componente accidentale $d_{\{x_e\}}$ di lunghezza proporzionale alla ricchezza percettiva dell'evento — i dettagli sensoriali, il rumore di fondo, le contingenze irrilevanti. Conservare $d_{\{x_e\}}$ costa esattamente $|d_{\{x_e\}}|$ bit di capacità di archiviazione per ogni evento, e questo costo cresce linearmente con il numero di eventi conservati.

Sia N il numero di eventi archiviati in un sistema con capacità C . La capacità media disponibile per evento è C/N . Se $|d_{\{x_e\}}| > C/N$ per l'evento tipico — se il residuo accidentale è più lungo della capacità media per evento — un archivio fedele è impossibile: il sistema non può conservare tutti gli eventi senza perdita. Poiché la capacità del sistema cognitivo è finita e il flusso di esperienze è continuo, questa condizione è inevitabilmente soddisfatta per un agente con storia sufficientemente lunga.

La conseguenza è che qualsiasi sistema cognitivo con storia lunga *deve* essere un compressore lossy. La domanda non è se comprimere con perdita — è quale struttura preservare e quale lasciar perdere. La risposta del sistema evolutivamente ottimizzato è: preservare $p_{\{x_e\}}$ (il modello predittivamente rilevante) e perdere $d_{\{x_e\}}$ (il residuo accidentale). La finzione è il risultato di questa scelta strutturale.

I.6 Tasso di distorsione e limite di Shannon-Kolmogorov

La teoria della compressione lossy fornisce un quadro formale preciso per analizzare il trade-off tra fedeltà e compressione. La *funzione di tasso-distorsione* $R(D)$ definisce la lunghezza minima della descrizione compressa compatibile con una distorsione media al più D , dove la distorsione è misurata da una funzione $d(x, \hat{x})$ che quantifica la differenza tra l'evento originale e il ricordo generato:

$$R(D) = \min_{\{P(\hat{x}|x) : \mathbb{E}[d(x, \hat{x})] \leq D\}} I(X; \hat{X})$$

dove $I(X; \hat{X})$ è l'informazione mutua tra la sorgente X e la rappresentazione compressa $\hat{X}$, e il minimo è preso su tutte le distribuzioni di codifica compatibili con il vincolo di distorsione. $R(D)$ è una funzione decrescente e convessa di D : minore è la distorsione tollerata, maggiore è la lunghezza descrittiva necessaria.

Per la memoria, la funzione di distorsione $d(x_e, \hat{x})$ non è la distanza euclidea tra eventi — non misura quanto il ricordo differisca dall'evento nella metrica percettiva. Misura quanto il ricordo differisca dall'evento nella metrica *predittiva*: $d(x_e, \hat{x})$ è grande se il ricordo $\hat{x}$ porta a previsioni sistematicamente sbagliate rispetto a quelle che l'evento originale x_e avrebbe giustificato. Questa scelta della funzione di distorsione è la formalizzazione dell'argomento del paragrafo precedente: il sistema ottimizza la coerenza predittiva, non la fedeltà percettiva.

Con questa scelta, $R(D)$ misura quanta informazione il sistema deve conservare per mantenere la distorsione predittiva entro D . Il limite di Shannon-Kolmogorov per la memoria cognitiva è:

$$KC(m_e) \geq R(D_{\text{pred}})$$

dove D_{pred} è la distorsione predittiva che il sistema tollera. La lunghezza della descrizione compressa m_e deve essere almeno $R(D_{\text{pred}})$ per garantire che i ricordi generati abbiano distorsione predittiva media al più D_{pred} . Questo è il lower bound sulla memoria: nessun sistema può ricordare con distorsione predittiva inferiore a D_{pred} conservando meno di $R(D_{\text{pred}})$ bit per evento.

I.7 La struttura della varianza di generazione: un modello parametrico

Si introduce ora un modello parametrico del processo generativo che permette di derivare previsioni verificabili sulla struttura della varianza mnestica.

Si modella il campionamento del ricordo come:

$$\hat{x} = f(m_e, \epsilon, \theta)$$

dove f è la funzione generativa, ϵ è rumore stocastico con distribuzione p_ϵ , e θ è lo stato corrente dell'agente. Si assume che f sia differenziabile rispetto a m_e e θ , e che p_ϵ abbia varianza σ^2 calibrata sulla ricchezza del residuo originale $d_{\{x_e\}}$: $\sigma^2 \propto KC(d_{\{x_e\}})$.

La varianza del ricordo rispetto alla variazione di θ — la *varianza contestuale* — è:

$$\text{Var}_\theta(\hat{x}) = (\partial f / \partial \theta)^2 \text{Var}(\theta) + \sigma^2$$

Il primo termine misura quanto il ricordo cambia al variare dello stato dell'agente — la componente sistematica della distorsione. Il secondo termine misura la varianza intrinseca del processo generativo — la componente stocastica.

La previsione chiave di questo modello è che la varianza contestuale $\text{Var}_\theta(\hat{x})$ cresce con $KC(d_{\{x_e\}})$: eventi con residuo accidentale più ricco producono ricordi con

varianza contestuale maggiore. Questo perché un residuo accidentale ricco significa che l'evento originale aveva molti dettagli non catturati dal modello — dettagli che il processo generativo deve "riempire" campionando da p_{ϵ} , e la varianza di questo campionamento è proporzionale alla quantità di informazione mancante.

La verifica empirica di questa previsione richiede di misurare la varianza inter-richiamo di ricordi dello stesso evento in condizioni contestuali diverse, e di correlarla con una stima della ricchezza percettiva dell'evento originale. Esperimenti di questo tipo — in cui lo stesso soggetto descrive lo stesso episodio autobiografico in momenti distanti e in stati emotivi diversi — mostrano una varianza inter-richiamo sistematicamente maggiore per eventi ricchi di dettagli sensoriali rispetto a eventi schematicamente strutturati, in accordo con la previsione.

La memoria come finzione

Capitolo II — Il riconsolidamento come riscrittura del programma generativo

II.1 Il paradosso del ricordo che cambia ricordando

Vi è un fenomeno nell'area del riconsolidamento mnestico che la psicologia ha impiegato decenni a riconoscere perché contraddiceva il modello dominante della memoria come archivio stabile: ogni volta che un ricordo viene *richiamato*, diventa temporaneamente vulnerabile alla modifica. Il ricordo riattivato ritorna in uno stato labile — neurobiologicamente, le sinapsi che lo sostengono vengono destabilizzate — e poi si *riconsolida*, ma in una forma che può differire dall'originale. Il richiamo non è soltanto lettura: è anche potenziale riscrittura.

Questo fenomeno — scoperto nella sua forma neurobiologica da Misanin, Miller e Lewis negli anni Sessanta su ratti, poi riscoperto e sistematizzato da Nader, Schafe e LeDoux nel 2000 — è stato interpretato come un difetto del sistema: la memoria è instabile, riscrivibile, fallibile. Ma nel quadro del modello generativo proposto nel Capitolo I, il riconsolidamento non è un difetto: è il meccanismo attraverso cui il sistema aggiorna il proprio programma generativo m_e ogni volta che genera un'istanza $\hat{x}$ — ogni volta che produce un ricordo — e confronta quella istanza con il contesto corrente θ . Il riconsolidamento è apprendimento continuo: la memoria si riscrive proprio perché ottimizza la coerenza predittiva, e la coerenza predittiva richiede che il modello si aggiorni ogni volta che nuova informazione è disponibile.

II.2 Il riconsolidamento come aggiornamento bayesiano

Si formalizza il riconsolidamento come aggiornamento bayesiano del modello generativo. Sia $m_e^{\{t\}}$ il modello mnestico per l'evento e al tempo t — la versione del programma generativo disponibile al momento del t -esimo richiamo. Al momento del richiamo, il sistema genera $\hat{x}^{\{t\}}$ dalla distribuzione $P(\hat{x} \mid m_e^{\{t\}}, c^{\{t\}}, \theta^{\{t\}})$.

Il ricordo generato $\hat{x}^{\{t\}}$ viene confrontato con il contesto corrente $\theta^{\{t\}}$, producendo un *segnale di errore predittivo*:

$$\delta^{\{t\}} = \log P_{\text{agent}}(\text{osservazioni}^{\{t\}} \mid \hat{x}^{\{t\}}, \text{azione}^{\{t\}}) - \mathbb{E}_{\hat{x}}[\log P_{\text{agent}}(\text{osservazioni}^{\{t\}} \mid \hat{x}, \text{azione}^{\{t\}})]$$

Questo segnale misura quanto il ricordo specifico generato ha meglio predetto le osservazioni rispetto alla media delle predizioni del modello. Se $\delta^{\{t\}} > 0$, il ricordo generato era più predittivamente accurato del previsto; se $\delta^{\{t\}} < 0$, era meno accurato.

Il riconsolidamento aggiorna il modello nella direzione del gradiente dell'errore predittivo:

$$m_e^{\{t+1\}} = m_e^{\{t\}} + \eta \cdot \nabla_{\{m_e\}} \delta^{\{t\}}$$

dove η è un tasso di apprendimento e $\nabla_{\{m_e\}}$ è il gradiente rispetto ai parametri del modello. Questo è un aggiornamento di gradiente stocastico sulla funzione obiettivo J definita nel Capitolo I: ogni richiamo fornisce un campione rumoroso del gradiente, e il modello converge — nel limite di molti richiami — verso il minimo di J .

La convergenza non è garantita in generale: se il tasso di apprendimento η è troppo alto, il sistema oscilla; se è troppo basso, l'aggiornamento è trascurabile. Ma il punto qualitativo è importante: il riconsolidamento non è un difetto accidentale del sistema mnestico — è il meccanismo di ottimizzazione on-line che aggiorna il programma generativo ogni volta che viene usato.

II.3 La deriva mnestica come traiettoria nel spazio dei modelli

Si considera ora la sequenza di modelli $m_e^{\{0\}}, m_e^{\{1\}}, \dots, m_e^{\{T\}}$ prodotta da T richiami successivi dell'evento e . Questa sequenza è una *traiettoria* nello spazio dei programmi generativi — una traiettoria guidata dal gradiente dell'errore predittivo, perturbata dal rumore del processo di campionamento.

La *deriva mnestica* è la distanza percorsa da questa traiettoria nel tempo:

$$\Delta_T = KC(m_e^{\{T\}} \mid m_e^{\{0\}})$$

la complessità condizionale del modello finale rispetto al modello iniziale — quanto informazione aggiuntiva è necessaria per descrivere $m_e^{\{T\}}$ una volta noto $m_e^{\{0\}}$. La deriva misura quanto il ricordo sia cambiato attraverso T richiami.

Si possono derivare alcune proprietà della deriva. La deriva è non-decrescente in T : ogni richiamo può solo aggiungere o mantenere la distanza dal modello originale, non ridurla (a meno di un richiamo che produca esattamente un aggiornamento che riporti il modello verso $m_e^{\{0\}}$, il che richiederebbe un errore predittivo di segno opposto agli aggiornamenti precedenti). La deriva è maggiore quando il contesto $\theta^{\{t\}}$ è sistematicamente diverso dal contesto originale dell'evento: richiami in stati emotivi molto diversi da quello in cui l'evento fu vissuto producono aggiornamenti sistematicamente distorti, perché l'errore predittivo viene calcolato rispetto a un contesto che non corrisponde all'evento originale.

La deriva è inoltre correlata con la frequenza dei richiami: eventi ricordati frequentemente derivano più rapidamente di eventi ricordati raramente, perché accumulano più aggiornamenti. Questo sembra controintuitivo — si ricordano meglio le cose che si ripetono — ma il modello distingue tra *accuratezza del modello corrente* (che può migliorare con il richiamo ripetuto, se gli aggiornamenti sono calibrati) e *fedeltà al modello originale* (che decresce con ogni aggiornamento). Un evento ricordato frequentemente è ricordato con alta coerenza predittiva nel contesto corrente, ma con bassa fedeltà all'esperienza originale.

II.4 Il teorema della deriva inevitabile

Enunciato. Sia $m_e^{\{t\}}$ la sequenza di modelli prodotta da richiami in contesti $\theta^{\{t\}}$ che variano nel tempo. Si assuma che le variazioni di θ abbiano varianza $\sigma_\theta^2 > 0$ e siano stazionarie. Allora la deriva attesa Δ_T cresce almeno come $\Omega(\log T)$ per $T \rightarrow \infty$.

Dimostrazione informale. Ogni aggiornamento introduce una perturbazione al modello di ampiezza almeno proporzionale a $\eta \cdot \sigma_\theta \cdot \|\nabla_{m_e} J\|$ — il prodotto del tasso di apprendimento, della variabilità contestuale e della norma del gradiente. Per un modello non al minimo assoluto di J , la norma del gradiente è positiva. La somma di T perturbazioni indipendenti con varianza positiva produce una deriva che cresce almeno come $\eta \cdot \sigma_\theta \cdot \Omega(\sqrt{T})$ nella norma del gradiente, che si traduce in una crescita almeno logaritmica di $KC(m_e^{\{T\}} \mid m_e^{\{0\}})$ per l'effetto della compressione sulla varianza. ■

Il corollario pratico è che nessun ricordo rimane stabile indefinitamente in un sistema che lo richiama. La stabilità assoluta richiede o l'assenza di richiami — che il ricordo rimanga "congelato" senza essere mai generato — o la costanza perfetta del contesto di richiamo, che è incompatibile con la condizione di un agente in un ambiente variabile. La deriva è una proprietà asintotica inevitabile del riconsolidamento: i ricordi di eventi lontani nel tempo, richiamati frequentemente in contesti variabili, sono sistematicamente i più distanti dall'evento originale.

II.5 Riconsolidamento e costruzione narrativa dell'identità

La conseguenza più rilevante del teorema di deriva inevitabile non è che i ricordi siano inaffidabili — questo è già noto — ma che la traiettoria della deriva non è casuale: è guidata dal gradiente dell'errore predittivo, il che significa che è guidata dalla struttura del contesto θ in cui i richiami avvengono.

Se i contesti di richiamo hanno una struttura comune — se l'agente richiama i propri ricordi prevalentemente in contesti emotivamente congruenti con un certo schema di sé — allora gli aggiornamenti di riconsolidamento saranno sistematicamente orientati verso quel schema. I ricordi derivano nella direzione del *modello narrativo di sé* che l'agente mantiene: si trasformano gradualmente in ricordi che sono coerenti con l'identità corrente dell'agente, non con l'esperienza originale.

Questo è il meccanismo formale della costruzione narrativa dell'identità autobiografica, che il Capitolo III analizzerà in dettaglio. Il punto che qui si vuole sottolineare è che la costruzione non è intenzionale — non è falsificazione deliberata — ma è una conseguenza strutturale dell'ottimizzazione on-line del modello generativo. L'identità narrativa emerge come *attrattore* della dinamica di riconsolidamento: i ricordi convergono verso configurazioni coerenti con il modello di sé, indipendentemente da ciò che l'agente intende deliberatamente.

II.6 Riconsolidamento selettivo: la stabilità dei ricordi emotivamente salienti

Non tutti i ricordi derivano alla stessa velocità. La neurobiologia del riconsolidamento ha mostrato che i ricordi con alta valenza emotiva — i ricordi di eventi fortemente positivi o negativi — sono più resistenti alla modifica attraverso il riconsolidamento rispetto ai ricordi emotivamente neutri. Questa asimmetria è stata interpretata come un meccanismo di protezione adattivo: i ricordi che portano informazione predittivamente rilevante sui pericoli o sulle opportunità dell'ambiente vengono stabilizzati per preservare quella informazione.

Nel quadro del modello generativo, questa asimmetria ha una spiegazione in termini di tasso di apprendimento η . Si propone che η non sia costante ma dipenda dalla valenza emotiva del ricordo richiamato:

$$\eta(e, t) = \eta_0 \cdot \exp(-\lambda \cdot |v(e)|)$$

dove $v(e)$ è la valenza emotiva dell'evento e e $\lambda > 0$ è un parametro di stabilizzazione. Per eventi con alta valenza $|v(e)| \gg 0$, il tasso di apprendimento è esponenzialmente ridotto: gli aggiornamenti di riconsolidamento sono piccoli, e la deriva è lenta. Per eventi emotivamente neutri, $\eta \approx \eta_0$ e la deriva è più rapida.

Questa formulazione predice che i ricordi traumatici — con valenza emotiva molto alta — siano i più stabili nel tempo e i più resistenti alla deriva narrativa. Predice anche, simmetricamente, che i ricordi di eventi emotivamente positivi siano più stabili

dei ricordi neutri, ma meno stabili dei ricordi traumatici. La letteratura sulla memoria flashbulb — i ricordi vividi di eventi fortemente emotivi — è in accordo qualitativo con questa previsione: tali ricordi sono più stabili nel tempo ma non necessariamente più accurati, il che si spiega con il fatto che la stabilità riduce la deriva ma non elimina la varianza di generazione σ^2 .

II.7 Il riconsolidamento come vulnerabilità terapeutica

Il riconsolidamento mnestico ha una conseguenza clinica che il modello formale rende perspicua. Se il riconsolidamento aggiorna $m_e^{\{t\}}$ in direzione del gradiente dell'errore predittivo, allora è possibile — in linea di principio — usare il riconsolidamento per *guidare* la deriva nella direzione voluta: indurre un richiamo in un contesto θ^* progettato per produrre aggiornamenti che spostano m_e verso una rappresentazione meno distressante dell'evento.

Questo è precisamente il meccanismo proposto dai protocolli di terapia che sfruttano il riconsolidamento — in particolare l'EMDR (Eye Movement Desensitization and Reprocessing) e alcune forme di terapia cognitivo-comportamentale focalizzata sul trauma. Il paziente riattiva il ricordo traumatico in un contesto di sicurezza ($\theta^* =$ contesto terapeutico), e il riconsolidamento aggiorna m_e nella direzione dell'errore predittivo calcolato rispetto a θ^* — errore che, in un contesto di sicurezza, è sistematicamente orientato verso la riduzione della valenza negativa.

Nel quadro formale, l'obiettivo terapeutico è modificare $v(e)$ — la valenza emotiva associata al ricordo — che a sua volta modifica $\eta(e, t)$: un ricordo traumatico che attraverso il riconsolidamento terapeutico vede ridotta la propria valenza diventa più plastico, più soggetto a deriva, e quindi più modificabile in direzioni adattive. La terapia, in questa lettura, non cancella il ricordo — $KC(m_e)$ rimane positiva — ma cambia la struttura del programma generativo in modo che le istanze generate siano meno distressanti e più compatibili con il funzionamento adattivo dell'agente.

Il modello suggerisce anche perché il riconsolidamento terapeutico abbia un finestra temporale critica: il ricordo è modificabile solo durante il periodo di labilità che segue il richiamo, tipicamente stimato in alcune ore. Questo corrisponde, nel modello, alla durata del periodo in cui il gradiente $\nabla_{\{m_e\}} J$ è accessibile per l'aggiornamento — dopo il quale il modello si "riconsolida" tornando a uno stato stabile e il tasso di apprendimento η torna al valore $\eta(e, t)$ dettato dalla valenza.

II.8 Contro il mito della memoria traumatica immutabile

Il modello di riconsolidamento permette di prendere posizione su un dibattito clinico importante: se i ricordi traumatici siano immutabili nel senso di conservare per sempre la propria struttura originale, o se siano soggetti a deriva come ogni altro ricordo.

La risposta nel quadro formale è che nessun ricordo è immutabile, ma i ricordi traumatici hanno un tasso di deriva molto più basso degli altri — per il meccanismo di stabilizzazione basato sulla valenza che la formula $\eta(e, t)$ cattura. La loro apparente

immutabilità è una deriva molto lenta, non l'assenza di deriva. Questa distinzione ha conseguenze cliniche: la lentezza della deriva non implica che il ricordo traumatico sia definitivamente fissato, ma che la finestra di riconsolidamento è più stretta e richiede interventi più precisi per essere sfruttata terapeutamente.

D'altra parte, il modello predice che interventi ripetuti di richiamo in contesti emotivamente distanti dal trauma — i cosiddetti *trigger* — possono paradossalmente *stabilizzare* il ricordo traumatico nella sua forma distressante: ogni richiamo in un contesto che produce errori predittivi negativi aggiorna m_e in direzione di una valenza ancora più negativa, aumentando $|v(e)|$ e riducendo ulteriormente η . Questo è il meccanismo formale della *sensibilizzazione* ai trigger: l'esposizione ripetuta non desensibilizza automaticamente, ma può iperstabilizzare il ricordo traumatico se avviene in contesti che confermano il modello negativo.

Capitolo III — La narrativa come codec: struttura dell'identità autobiografica

III.1 Il problema della coerenza autobiografica

Un agente cognitivo che attraversasse la propria vita accumulando ricordi come istanze indipendenti — una collezione di scene senza relazioni causali tra loro — non disporrebbe di ciò che si chiama identità. L'identità richiede che i ricordi siano organizzati in una struttura che li connette: che ogni evento sia inserito in una catena causale che risale al passato e si proietta verso il futuro, che il protagonista di quelle scene sia riconoscibile come lo stesso agente attraverso il tempo, che le proprie azioni abbiano motivazioni coerenti con un carattere che persiste. Questa struttura è la *narrativa autobiografica*.

Il paradosso che il Capitolo I ha aperto è che la narrativa autobiografica non può essere la registrazione fedele dell'esperienza: il processo generativo introduce varianza, il riconsolidamento introduce deriva, e la capacità di archiviazione è finita. La narrativa deve essere costruita, non recuperata. La domanda che questo capitolo affronta è: con quale struttura formale viene costruita, e quali sono le conseguenze di quella struttura per l'accuratezza del ricordo e per la coerenza dell'identità?

La tesi è che la narrativa autobiografica è un *codec*: un sistema di codifica e decodifica che comprime l'esperienza grezza in una rappresentazione strutturata, al costo di una distorsione sistematica che non è casuale ma ha la forma specifica impressa dalla struttura del codec. Comprendere il codec è comprendere la forma della distorsione — e quindi prevedere dove e come la memoria mente, e perché mente sempre nella stessa direzione.

III.2 La struttura formale del codec narrativo

Un codec, nel senso della teoria dell'informazione, è una coppia (E, D) dove E è un codificatore — una funzione che mappa la sorgente nella rappresentazione compressa — e D è un decodificatore — una funzione che ricostruisce un'approssimazione della sorgente dalla rappresentazione compressa. Per un codec lossy, $D(E(x)) \neq x$ in generale: la ricostruzione è approssimata.

Il codec narrativo è la coppia $(E_{\text{narr}}, D_{\text{narr}})$ dove:

$$E_{\text{narr}} : \Omega^T \rightarrow \mathcal{N}$$

mappa una sequenza di T eventi esperienziali $x_{\{e_1\}}, \dots, x_{\{e_T\}}$ nello spazio delle narrazioni $\mathcal{N}$ — strutture con agente, motivazioni, causalità, temporalità e arco — e

$$D_{\text{narr}} : \mathcal{N} \times \mathcal{Q} \rightarrow \Omega$$

decodifica, data una narrazione $n \in \mathcal{N}$ e una query $q \in \mathcal{Q}$ (una richiesta di ricordare un evento specifico), un'approssimazione $\hat{x}$ dell'evento originale.

La compressione operata da E_{narr} è strutturata: non è una compressione generica che elimina i bit meno informativi, ma una compressione che preserva specificamente la struttura causale, la coerenza del carattere del protagonista, la progressione temporale. Formalmente, E_{narr} è ottimizzato per minimizzare la lunghezza della rappresentazione compressa $KC(n)$ compatibilmente con la conservazione della struttura causale — la componente del segnale che spiega perché gli eventi si sono succeduti nel modo in cui si sono succeduti.

III.3 La grammatica della narrativa come sistema di vincoli

La struttura del codec narrativo può essere specificata attraverso un sistema di vincoli che la narrazione deve soddisfare. Si identificano quattro vincoli principali.

Il primo è il *vincolo causale*: ogni evento deve essere causalmente connesso agli eventi precedenti e successivi. Una narrazione in cui gli eventi sono casualmente disconnessi non è riconoscibile come narrazione — è un elenco. La connessione causale è la primitiva fondamentale del codec narrativo, e la sua conservazione è il requisito che il codificatore E_{narr} non può violare senza produrre una non-narrazione.

Il secondo è il *vincolo di coerenza del carattere*: le azioni del protagonista devono essere spiegabili da un insieme stabile di tratti, desideri e credenze che persistono attraverso il tempo. Un protagonista che cambia carattere in modo imprevedibile e non motivato non è un personaggio narrativo — è un artefatto. La coerenza del carattere è il vincolo che impone che il modello del sé dell'agente sia comprimibile: che esista un programma breve $p_{\{\text{sé}\}}$ tale che le azioni del protagonista siano generate da $p_{\{\text{sé}\}}$ applicato alle circostanze.

Il terzo è il *vincolo di completezza causale*: se un evento ha conseguenze, quelle conseguenze devono apparire nella narrazione. Gli eventi senza conseguenze visibili tendono a essere eliminati dal codec — vengono trattati come rumore e compressi via. La completezza causale è il meccanismo attraverso cui la narrativa seleziona quali eventi preservare: quelli che entrano nella catena causale che spiega lo stato attuale dell'agente.

Il quarto è il *vincolo di arco*: la narrazione deve avere una struttura orientata — deve condurre da uno stato iniziale a uno stato corrente attraverso una trasformazione riconoscibile. L'arco narrativo è la struttura globale che dà senso alle catene causali locali: perché ciò che è accaduto ha reso l'agente ciò che è.

III.4 La distorsione sistematica indotta dal codec

Ogni vincolo del codec narrativo introduce una distorsione sistematica nel processo di ricostruzione. Le distorsioni non sono casuali — hanno direzione e struttura — e sono quindi, in linea di principio, prevedibili.

Il vincolo causale produce la *distorsione teleologica*: gli eventi vengono ricostruiti come più causalmente coerenti di quanto fossero nella loro struttura originale. La memoria riempie le lacune causali — i momenti in cui la catena causale era incerta o assente — con ricostruzioni che rispettano la causalità. Gli studi di Michotte sulla percezione causale, e poi la vasta letteratura sullo "hindsight bias", documentano questa distorsione: gli eventi passati vengono ricordati come più prevedibili di quanto fossero, perché il codec narrativo ha già inserito la connessione causale che non era visibile al momento.

La distorsione teleologica può essere formalizzata. Sia $C(x_e)$ la "causalità percepita" dell'evento e nella narrazione — una misura di quanto l'evento sia causalmente connesso agli eventi adiacenti nella rappresentazione compressa. La distorsione del vincolo causale è:

$$d_{\text{causal}}(x_e, \hat{x}_e) = C(\hat{x}_e) - C(x_e) \geq 0$$

la ricostruzione ha sempre causalità percepita almeno uguale all'originale — il codec non può ridurla senza violare il vincolo. La distorsione è sistematicamente positiva: i ricordi sembrano sempre più causalmente coerenti dell'esperienza originale.

Il vincolo di coerenza del carattere produce la *distorsione di consistenza*: le azioni e le credenze passate vengono ricostruite come più coerenti con il carattere corrente dell'agente di quanto fossero. Uno studio classico di Ross e Conway (1984) mostrò che i soggetti ricordano le proprie credenze passate come più simili alle credenze attuali di quanto registrazioni indipendenti documentino: il codec narrativo riscrive le credenze passate per mantenerle coerenti con il modello del sé corrente.

La distorsione di consistenza è proporzionale alla distanza tra il carattere corrente $p_{\{\text{sé}\}^{\{t\}}}$ e il carattere al momento dell'evento $p_{\{\text{sé}\}^{\{0\}}}$:

$$d_{\text{cons}}(\hat{x}_e, x_e) \propto KC(p_{\{\text{sé}\}^{\{t\}}} \mid p_{\{\text{sé}\}^{\{0\}}})$$

più il carattere è cambiato nel tempo, più le credenze e le azioni passate vengono distorte nella direzione del carattere corrente. Questa è la formalizzazione della *consistenza retrospettiva*: il sé narrativo di oggi riscrive il sé di ieri per mantenere l'illusione di una continuità che la deriva temporale ha in parte eroso.

Il vincolo di completezza causale produce la *distorsione di omissione*: gli eventi senza conseguenze visibili vengono sistematicamente eliminati o compressi in modo più estremo degli eventi causalmente rilevanti. La vita quotidiana è prevalentemente composta di eventi routine senza conseguenze narrative — mattine identiche, conversazioni banali, spostamenti ripetuti. Questi vengono compressi in schemi generici ("di solito facevo X") mentre gli eventi causalmente rilevanti vengono conservati con maggiore dettaglio.

Il vincolo di arco produce la *distorsione di significato*: gli eventi vengono ricostruiti come più significativi — più orientati verso il tema del proprio arco di vita — di quanto fossero nella loro struttura originale. Gli studi sulla narrazione di vita mostrano che i soggetti organizzano la propria storia autobiografica secondo pochi temi dominanti — McAdams chiama questi temi "immagini nucleari" — e che gli eventi vengono ricostruiti come manifestazioni di quei temi anche quando la loro connessione con il tema era tenue o assente nell'esperienza originale.

III.5 La funzione di tasso-distorsione del codec narrativo

Il codec narrativo ha una propria funzione di tasso-distorsione che opera in parallelo con quella definita nel Capitolo I. Per il codec narrativo, la distorsione non è misurata nella metrica predittiva bensì nella metrica della *coerenza narrativa*: $d_{\text{narr}}(n, n')$ misura quanto le narrazioni n e n' differiscano nella struttura causale, nella coerenza del carattere, nella completezza causale e nell'arco.

La funzione di tasso-distorsione narrativa $R_{\text{narr}}(D_{\text{narr}})$ definisce la lunghezza minima della narrazione compressa compatibile con una distorsione narrativa media al più D_{narr} :

$$R_{\text{narr}}(D_{\text{narr}}) = \min_{\{P(n|x^T) : \mathbb{E}[d_{\text{narr}}] \leq D_{\text{narr}}\}} I(X^T; N)$$

dove X^T è la sequenza di eventi esperienziali e N è la narrazione compressa. R_{narr} misura quanta informazione la narrazione deve conservare per mantenere la distorsione narrativa entro D_{narr} .

La relazione tra R_{narr} e la funzione di tasso-distorsione predittiva R del Capitolo I è non ovvia: le due funzioni misurano distorsioni diverse — predittiva e narrativa — e i loro minimi non coincidono in generale. Un evento può essere predittivamente irrilevante (bassa distorsione predittiva se omissa) ma narrativamente cruciale (alta distorsione narrativa se omissa), e viceversa. La tensione tra le due funzioni obiettivo

è la fonte delle distorsioni più interessanti della memoria autobiografica: i ricordi che vengono conservati in modo vivido ma distorto sono spesso quelli in cui le due metriche sono in conflitto.

III.6 La narrativa del sé come programma generativo di ordine superiore

Si introduce ora la nozione centrale di questo capitolo: la narrativa autobiografica non è soltanto una struttura di archiviazione, ma un *programma generativo di ordine superiore* — un metaprogramma $p_{\{sé\}}$ che genera non solo i ricordi specifici degli eventi passati, ma anche le aspettative sugli eventi futuri, le interpretazioni degli eventi presenti, e le reazioni emotive agli stimoli ambientali.

Formalmente, $p_{\{sé\}}$ è un programma che mappa contesti c in distribuzioni sugli eventi attesi:

$$p_{\{sé\}}(c) = P_{\{sé\}}(\text{evento} \mid \text{contesto } c)$$

Questa distribuzione è la prior dell'agente su ciò che può accadere nel contesto c , modellata dal proprio carattere narrativo. Gli eventi che si verificano vengono confrontati con $p_{\{sé\}}(c)$ — producono un errore predittivo $\delta_{sé}$ — e questo errore alimenta l'aggiornamento di riconsolidamento del Capitolo II.

Il punto cruciale è che $p_{\{sé\}}$ è il programma generativo che vincola la forma delle distorsioni narrative: le distorsioni di consistenza, di completezza causale e di arco sono tutte proiezioni di $p_{\{sé\}}$ sull'esperienza passata. Il sé narrativo non è soltanto il protagonista della storia: è il programma che genera la storia. Comprendere la struttura di $p_{\{sé\}}$ significa comprendere la forma delle distorsioni — dove e come la memoria autobiografica mente, e perché le sue menzogne hanno sempre la stessa sagoma.

III.7 La sofisticazione del sé narrativo e la patologia delle narrative rigide

La sofisticazione di $p_{\{sé\}}$ — la lunghezza della parte strutturalmente necessaria del programma del sé — misura la ricchezza della narrativa autobiografica: quanto articolata, sfumata e internamente differenziata sia la struttura del sé. Un $p_{\{sé\}}$ con alta sofisticazione genera una narrativa ricca di distinzioni interne, capace di integrare esperienze eterogenee senza ridurle a schemi uniformi.

Un $p_{\{sé\}}$ con bassa sofisticazione genera una narrativa rigida: pochi temi dominanti, distorsioni di consistenza elevate, scarsa capacità di integrare esperienze che contraddicono il modello corrente. Le patologie della narrativa autobiografica — la ruminazione, il pensiero rigido, certe forme di depressione e di disturbo di personalità — si manifestano formalmente come sistemi con $p_{\{sé\}}$ a bassa sofisticazione: il

programma del sé ha poche primitive, e ogni esperienza viene compressa in uno dei pochi schemi disponibili, con distorsioni di consistenza molto elevate.

Si formalizza questo come segue. La *rigidità narrativa* di un agente è:

$$\text{rig} = 1 / \text{soph}(p_{\{\text{sé}\}})$$

l'inverso della sofisticazione del programma del sé. Alta rigidità corrisponde a bassa sofisticazione: il sé narrativo ha poche strutture interne e alta compressione. La rigidità predice l'entità delle distorsioni di consistenza:

$$d_{\text{cons}}(\hat{x}_e, x_e) \propto \text{rig} \cdot \text{KC}(p_{\{\text{sé}\}}^{\{(t)\}} \mid p_{\{\text{sé}\}}^{\{(0)\}})$$

un agente con alta rigidità distorce i ricordi passati più intensamente verso il modello corrente, perché il suo programma del sé ha poche primitive e deve comprimere ogni esperienza in uno schema preesistente.

La terapia narrativa — nel senso ampio che comprende sia la psicoterapia tradizionale che le pratiche di revisione autobiografica — può essere interpretata formalmente come un processo di aumento della sofisticazione di $p_{\{\text{sé}\}}$: l'agente acquisisce nuove primitive narrative — distinzioni concettuali, prospettive alternative, strutture causali più articolate — che abbassano rig e riducono le distorsioni di consistenza. Il guadagno terapeutico è misurabile, in linea di principio, come aumento di $\text{soph}(p_{\{\text{sé}\}})$ e riduzione della rigidità.

III.8 Il codec narrativo come tecnologia evolutiva

Si conclude il capitolo con una considerazione che il modello formale suggerisce e che la letteratura evolutiva non ha esplorato sistematicamente. Il codec narrativo — la struttura $(E_{\text{narr}}, D_{\text{narr}})$ che organizza l'esperienza in narrazione — non è una proprietà universale di qualsiasi sistema cognitivo: è una tecnologia specifica che richiede, per funzionare, un sistema di rappresentazione con le primitive appropriate — la causalità, la temporalità, l'intenzionalità, il sé come agente.

Questa tecnologia ha un costo: le distorsioni sistematiche che i vincoli narrativi introducono. Ha anche un beneficio: la coerenza predittiva che la struttura narrativa conferisce all'agente — la capacità di anticipare eventi in domini causalmente strutturati, di pianificare su orizzonti lunghi, di coordinare le proprie azioni con quelle di altri agenti che condividono la stessa grammatica narrativa.

Il trade-off può essere formalizzato. Il valore del codec narrativo per l'agente è la differenza tra la riduzione di distorsione predittiva che esso produce — rispetto a un sistema senza struttura narrativa — e il costo delle distorsioni narrative che introduce:

$$V(\text{codec}) = \Delta R_{\text{pred}}(D_{\text{pred}}) - \Delta R_{\text{narr}}(D_{\text{narr}})$$

dove ΔR_{pred} è il guadagno di efficienza predittiva e ΔR_{narr} è il costo aggiuntivo di distorsione narrativa. Il codec è evolutivamente vantaggioso quando $V > 0$: quando il guadagno predittivo supera il costo distorsivo.

In ambienti altamente strutturati causalmente — ambienti in cui le relazioni tra eventi sono stabili, in cui il carattere degli agenti è predittivo del loro comportamento, in cui gli archi narrativi si ripetono — il guadagno predittivo è elevato e $V \gg 0$. In ambienti caotici o in rapido cambiamento — in cui le strutture causali si rompono, i caratteri cambiano imprevedibilmente, gli archi narrativi si interrompono — il guadagno predittivo si riduce mentre il costo distorsivo rimane, e V può diventare negativo: il codec narrativo diventa una fonte di errore piuttosto che di precisione.

Questo suggerisce una previsione: gli agenti in ambienti molto instabili dovrebbero beneficiare di sistemi di rappresentazione con narrativa più debole — meno vincoli di causalità e di coerenza del carattere — rispetto a agenti in ambienti stabili. La neurobiologia delle popolazioni umane in ambienti ecologici diversi non ha ancora testato questa previsione in modo sistematico, ma i dati sulla flessibilità cognitiva e sull'adattamento a contesti di alta incertezza sono qualitativamente consistenti con essa.

Capitolo IV — La confabulazione come inferenza bayesiana

IV.1 La confabulazione e il suo paradosso

La confabulazione — la produzione spontanea di ricordi falsi senza intenzione di ingannare, tipicamente osservata in pazienti con lesioni al lobo frontale o con sindrome di Korsakoff — è il fenomeno che più direttamente sfida il modello della memoria come archivio. Il paziente confabulante non sa di mentire: produce con piena convinzione soggettiva ricordi di eventi che non sono mai accaduti, che contraddicono la cronologia verificabile, che mescolano elementi di episodi reali in combinazioni impossibili. Il confabulante non è un impostore: è un sistema generativo che campiona dalla propria distribuzione a priori in assenza dell'informazione — la traccia mnestica originale — che normalmente vincola il campionamento.

Questa è la tesi di questo capitolo: la confabulazione non è un difetto del processo mnestico che si aggiunge alla sua struttura normale — è la struttura normale del processo mnestico, resa visibile dalla rimozione del vincolo esterno che normalmente la disciplina. Nella memoria ordinaria, il campionamento dalla distribuzione generativa è vincolato dalla traccia m_e : il modello deve produrre istanze compatibili con ciò che è stato registrato. Quando la traccia è assente o degradata — per lesione, per intossicazione, per semplice distanza temporale — il vincolo viene meno, e il sistema continua a campionare dalla propria distribuzione a priori. La confabulazione è generazione senza vincolo di verosimiglianza.

IV.2 Il modello bayesiano della memoria come inferenza

Si formalizza la memoria come un problema di inferenza bayesiana. L'agente deve inferire il contenuto dell'evento passato e — la variabile nascosta — dati tre insiemi di informazione: la cue di recupero c , il modello compresso m_e (o ciò che ne rimane dopo la degradazione), e le proprie credenze a priori $p(e)$ sull'evento — la distribuzione che il programma del sé $p_{\{sé\}}$ induce sugli eventi probabili nel contesto dell'evento.

La distribuzione a posteriori sull'evento è data dalla regola di Bayes:

$$p(e \mid c, m_e, p_{\{sé\}}) \propto p(c \mid e) \cdot p(m_e \mid e) \cdot p(e \mid p_{\{sé\}})$$

I tre fattori sono il *likelihood della cue* $p(c \mid e)$ — quanto la cue è compatibile con l'evento e —, il *likelihood della traccia* $p(m_e \mid e)$ — quanto la traccia compressa è compatibile con e — e la *prior* $p(e \mid p_{\{sé\}})$ — quanto e è a priori probabile dato il modello del sé dell'agente.

Il ricordo generato $\hat{x}$ è un campione dall'istanza dell'evento e^* che massimizza questa distribuzione a posteriori, o equivalentemente che minimizza l'energia libera variazionale:

$$\begin{aligned} e^* &= \operatorname{argmax}_e p(e \mid c, m_e, p_{\{sé\}}) \\ &= \operatorname{argmin}_e [-\log p(c \mid e) - \log p(m_e \mid e) - \log p(e \mid p_{\{sé\}})] \end{aligned}$$

I tre termini corrispondono a tre sorgenti di vincolo: la cue, la traccia mnestica e la prior narrativa. Il ricordo è l'istanza dell'evento che bilancia ottimamente questi tre vincoli.

IV.3 La confabulazione come dominanza della prior

Nel caso normale, la traccia m_e è sufficientemente informativa da vincolare fortemente la distribuzione a posteriori: il termine $-\log p(m_e \mid e)$ penalizza pesantemente le istanze incompatibili con la traccia, e la prior ha un effetto limitato. Il ricordo generato è vicino all'evento originale — la varianza di generazione è dominata dalla traccia, non dalla prior.

Quando la traccia è degradata — il termine $-\log p(m_e \mid e)$ è piatto, cioè la traccia non discrimina tra eventi compatibili e incompatibili — la distribuzione a posteriori è dominata dalla prior $p(e \mid p_{\{sé\}})$. Il sistema campiona dall'evento più a priori probabile dato il contesto e il carattere narrativo del soggetto. Questo campione è plausibile, coerente con il carattere del soggetto, contestualmente appropriato — ma non necessariamente corrispondente a ciò che è accaduto.

Questa è la confabulazione: campionamento dalla prior in assenza di vincolo di verosimiglianza. La prior è il programma del sé $p_{\{sé\}}$ — la distribuzione che il

modello narrativo del soggetto induce sugli eventi probabili. I confabulanti non generano eventi arbitrari: generano eventi plausibili per la propria narrativa autobiografica, eventi che *avrebbero potuto* accadere, eventi compatibili con il proprio carattere e con le proprie routines. La struttura dei contenuti confabulati riflette direttamente la struttura di $p_{\{sé\}}$.

IV.4 Una tassonomia formale degli errori mnestici

Il modello bayesiano permette di costruire una tassonomia degli errori mnestici basata su quale dei tre fattori nella distribuzione a posteriori è distorto. Questa tassonomia non è fenomenologica — non classifica gli errori per la loro apparenza — ma eziologica: classifica gli errori per la loro sorgente computazionale.

Gli *errori di cue* emergono quando $p(c | e)$ è mal calibrato: la cue attiva eventi errati perché la struttura dell'indizio è simile a quella di un evento diverso da quello target. I classici esperimenti di Tulving e Thomson sulla specificità di codifica documentano questo tipo di errore: una parola appresa in un contesto semantico viene difficilmente recuperata da una cue di contesto diverso, anche se la parola è "in memoria" nel senso che sarebbe riconosciuta se presentata. La distorsione non è nella traccia ma nella mappa tra cue ed evento.

Gli *errori di traccia* emergono quando $p(m_e | e)$ è degradato: la traccia compressa ha perso informazione sufficiente a vincolare il campionamento. Gli errori di questo tipo crescono con il tempo dall'evento originale — la traccia si degrada — e con la frequenza dei richiami — il riconsolidamento introduce deriva. I falsi ricordi di dettagli periferici di un evento ben ricordato nella sua struttura principale appartengono tipicamente a questa categoria: la traccia del modello m_e è intatta, ma il residuo d_e è andato perduto, e il sistema campiona d_e dalla prior anziché dalla traccia.

Gli *errori di prior* emergono quando $p(e | p_{\{sé\}})$ è mal calibrato: la distribuzione a priori dell'agente sugli eventi probabili è distorta, e questa distorsione contamina la distribuzione a posteriori anche quando la traccia è presente. Gli errori di prior sono sistematici e ripetibili — lo stesso soggetto fa lo stesso tipo di errore in situazioni diverse — perché riflettono la struttura fissa di $p_{\{sé\}}$. Gli errori di stereotipo — ricordare eventi in accordo con aspettative culturali anziché con l'esperienza effettiva — sono errori di prior: la distribuzione a priori è dominata dalla struttura categoriale anziché dall'istanza specifica.

IV.5 La formula di Bayes applicata alla sindrome di Korsakoff

La sindrome di Korsakoff — tipicamente causata da deficit di tiamina associato ad alcolismo cronico — produce confabulazione massiva in cui i pazienti hanno ricordi molto vividi di eventi recenti che non sono mai accaduti, spesso mescolando elementi di episodi passati reali in combinazioni temporalmente impossibili.

Nel modello bayesiano, la sindrome di Korsakoff corrisponde a una lesione selettiva del meccanismo che calcola $p(m_e | e)$: le tracce mnestiche degli eventi recenti non

vengono consolidate nel passaggio dall'ippocampo alla corteccia (il meccanismo anatomico-funzionale del consolidamento sistemico), e il termine di verosimiglianza della traccia è uniformemente piatto per gli eventi recenti. La distribuzione a posteriori è dunque dominata dalla prior $p(e \mid p_{\{sé\}})$, e il sistema genera eventi plausibili per il proprio carattere autobiografico.

La previsione formale è che i contenuti delle confabulazioni di Korsakoff debbano riflettere statisticamente la struttura di $p_{\{sé\}}$ del paziente: devono essere eventi del tipo che il paziente ha frequentemente esperito nella propria vita, in contesti che il paziente ha frequentemente abitato, con personaggi che il paziente ha frequentemente incontrato. La letteratura clinica sulla sindrome di Korsakoff è qualitativamente consistente con questa previsione: i pazienti confabulano eventi di vita quotidiana plausibili, non eventi fantastici o bizzarri. Berlyne (1972) e Kopelman (1987) documentano che i confabulanti raramente producono contenuti che contraddicono il proprio schema autobiografico — producono istanze della prior, non deviazioni da essa.

IV.6 La quantificazione della plausibilità confabulatoria

Si può derivare una misura formale della *plausibilità confabulatoria* di un contenuto mnestico — la probabilità che quel contenuto sia il prodotto di confabulazione anziché di un ricordo genuino. Si definisce:

$$\text{ConfScore}(\hat{x}) = \log p(\hat{x} \mid p_{\{sé\}}) - \log p(\hat{x} \mid m_e)$$

il logaritmo del rapporto tra la probabilità dell'istanza $\hat{x}$ sotto la prior e la sua probabilità sotto la traccia mnestica. Un ConfScore elevato indica che $\hat{x}$ è molto più probabile sotto la prior che sotto la traccia — è il tipo di ricordo che il sistema avrebbe generato anche in assenza della traccia. Un ConfScore basso indica che $\hat{x}$ è specificamente supportato dalla traccia — è improbabile sotto la prior e quindi difficilmente confabulabile.

Questa misura ha un'implicazione diagnostica: i contenuti mnestici con ConfScore elevato sono più suscettibili di distorsione da confabulazione — sono quelli per cui la prior e la traccia concordano, rendendo impossibile distinguere campionamento dalla traccia da campionamento dalla prior. I contenuti con ConfScore basso — eventi inattesi, contronormativi, specifici — sono meno suscettibili di distorsione: la loro bassa probabilità sotto la prior li rende tracciabili proprio perché il sistema non li avrebbe generati autonomamente.

Questa previsione è verificabile e ha conseguenze per la pratica testimoniale. I dettagli di un evento che sono altamente tipici — il colore delle pareti di un ufficio standard, l'ordine degli eventi in una routine familiare — hanno ConfScore elevato e sono quindi più suscettibili di confabulazione. I dettagli atipici — un particolare insolito, un evento contronormativo — hanno ConfScore basso e sono più affidabili come indicatori di memoria genuina, quando presenti. La letteratura sulla valutazione della

testimonianza oculare è qualitativamente consistente con questa previsione, sebbene raramente la formalizzi in termini bayesiani.

IV.7 L'interferenza tra prior e traccia: il falso ricordo come caso intermedio

Tra la memoria genuina — dominata dalla traccia — e la confabulazione — dominata dalla prior — esiste una vasta regione di casi intermedi in cui la traccia e la prior sono entrambe attive ma in conflitto. Questi casi intermedi producono i *falsi ricordi* nel senso di Loftus: ricordi che sono parzialmente radicati in un evento reale ma che contengono elementi generati dalla prior.

Si modella questo caso come interferenza tra i due termini nella distribuzione a posteriori. Sia e_true l'evento originale e sia e_prior l'evento più probabile sotto la prior per quella cue. La distribuzione a posteriori ha due modi — due massimi locali — uno intorno a e_true e uno intorno a e_prior . Il ricordo generato campiona da questa distribuzione bimodale, e può cadere nell'uno o nell'altro bacino, o produrre un'istanza che è una interpolazione tra i due.

Formalmente, il ricordo generato in presenza di interferenza è:

$$\hat{x} = \alpha \cdot x_{\{e_true\}} + (1-\alpha) \cdot x_{\{e_prior\}} + \varepsilon$$

dove $\alpha \in [0,1]$ è il peso relativo della traccia rispetto alla prior nella distribuzione a posteriori, e ε è rumore di campionamento. Il peso α è:

$$\alpha = \sigma(\log p(m_e \mid e_true) - \log p(e_prior \mid p_{\{s\}}) + \beta)$$

dove σ è la funzione sigmoide e β è un bias che dipende dallo stato dell'agente. α cresce con la forza della traccia ($p(m_e \mid e_true)$ alta) e decresce con la dominanza della prior ($p(e_prior \mid p_{\{s\}})$ alta).

Questa formula predice la struttura dei falsi ricordi: il ricordo falso è un'interpolazione tra l'evento originale e l'evento più a priori probabile nel contesto dato. I dettagli del ricordo falso che si discostano dall'originale devono essere nella direzione dell'evento più plausibile sotto la prior — non deviazioni arbitrarie, ma deviazioni sistematicamente orientate verso la plausibilità narrativa. La vasta letteratura sui falsi ricordi in contesti di suggestione sperimentale — a partire dai lavori di Loftus sulla domanda che introduce informazioni fuorvianti — documenta esattamente questa struttura: i falsi dettagli introdotti sono sempre plausibili nel contesto dell'evento, mai bizzarri o incongruenti.

IV.8 La prior mal calibrata: pregiudizio, stereotipo e trauma

Il modello bayesiano permette di analizzare formalmente tre fenomeni che la letteratura tratta separatamente ma che condividono la stessa struttura

computazionale: il pregiudizio implicito, la memoria stereotipica e la memoria traumatica.

Il *pregiudizio implicito* corrisponde a una prior $p(e | p_{\{sé\}})$ sistematicamente distorta nella direzione di aspettative categoriali. Un agente con pregiudizi razziali ha una prior che assegna alta probabilità a eventi negativi associati a certi gruppi, indipendentemente dall'evidenza specifica dell'evento. Quando la traccia mnestica dell'evento è parzialmente degradata — come accade per eventi in cui il soggetto non era molto attento — la prior domina, e il ricordo generato è distorto nella direzione del pregiudizio. Il meccanismo non è un atto deliberato di distorsione: è inferenza bayesiana con prior mal calibrata.

La *memoria stereotipica* è la versione meno intensa dello stesso fenomeno: le categorie culturali — i prototipi, gli script, gli schemi — costituiscono una prior molto forte che sovrascrive i dettagli idiosincratici degli eventi specifici quando la traccia è debole. Il ricordo di un evento che si discosta dallo script culturale atteso — per esempio, un evento di vita quotidiana che si svolge in modo insolito — tende a essere ricostruito nella direzione dello script, correggendo la deviazione insolita verso la norma. Questo è documentato sistematicamente dalla letteratura sulle script theory (Schank e Abelson, 1977) e dai successivi lavori sull'influenza degli schemi sulla memoria.

La *memoria traumatica* è il caso opposto: una prior dominata dall'evento traumatico, che sovrasaturava il sistema di allerta dell'agente con alta valenza negativa. In questo caso, la prior $p(e | p_{\{sé\}})$ nel contesto di cues associate al trauma assegna altissima probabilità all'evento traumatico, e anche tracce mnestiche parziali di eventi ordinari vengono interpretate come istanze del trauma. Il meccanismo del flashback — il ricordo intrusivo del trauma in risposta a cues parzialmente associate — è confabulazione in direzione del trauma: la prior traumatica domina anche quando la traccia non supporta il recupero dell'evento traumatico.

IV.9 Il valore epistemico della confabulazione

Si chiude il capitolo con una considerazione che il titolo della monografia rende necessaria: la confabulazione — la finzione mnestica — ha un valore epistemico? Il sistema che genera ricordi plausibili invece di ricordi accurati produce qualcosa di utile?

La risposta, nel quadro formale, è affermativa in senso limitato e precisamente delimitato. La prior $p(e | p_{\{sé\}})$ è una distribuzione che codifica la struttura statistica dell'esperienza passata dell'agente — le regolarità del proprio ambiente, le proprie routines, la struttura causale del proprio mondo sociale. Campionare dalla prior produce istanze che sono statisticamente tipiche di quell'esperienza, e quindi potenzialmente predittivamente rilevanti anche in assenza di traccia specifica. La confabulazione è, in questo senso, il sistema che continua a generare previsioni plausibili anche quando le tracce specifiche sono assenti: è un meccanismo di degrado graduale anziché di fallimento categorico.

Ma il valore epistemico è strettamente limitato dai contesti in cui la tipicità statistica è una buona approssimazione dell'accuratezza specifica — contesti stabili, routinari, altamente strutturati. Nei contesti in cui ciò che conta è precisamente la specificità dell'evento — il testimone di un delitto, il paziente che riferisce i propri sintomi, il soggetto sperimentale che deve riportare la propria esperienza — la confabulazione produce un output sistematicamente fuorviante. Il valore epistemico della finzione mnestica è inversamente proporzionale all'importanza della specificità: massimo dove la specificità è irrilevante, nullo dove la specificità è tutto ciò che conta.

Capitolo V — La memoria collettiva come sistema distribuito di aggiornamento narrativo

V.1 Dalla memoria individuale alla memoria collettiva: il problema del salto di scala

Il modello sviluppato nei capitoli precedenti descrive un sistema cognitivo individuale: un agente con un programma del sé $p_{\text{sé}}$, un insieme di modelli mnestici m_e , un processo di riconsolidamento che aggiorna quei modelli ad ogni richiamo. Quando si passa dalla memoria individuale alla memoria collettiva — la memoria condivisa di un gruppo, una comunità, una nazione — si incontra una discontinuità che non è soltanto di scala ma di struttura.

La memoria collettiva non è la somma delle memorie individuali. Non è nemmeno la loro intersezione — l'insieme di ciò che tutti i membri del gruppo ricordano allo stesso modo. È una struttura emergente: un sistema di rappresentazioni distribuite su una rete di agenti, stabilizzato attraverso pratiche sociali di richiamo — rituali, narrazioni condivise, monumenti, archivi, conversazioni — che producono sincronizzazione tra i modelli mnestici individuali e aggiornamento collettivo attraverso meccanismi analoghi al riconsolidamento, ma operanti sull'intera rete.

La tesi di questo capitolo è che la memoria collettiva è un sistema generativo distribuito con le stesse proprietà strutturali della memoria individuale — compressione lossy, riconsolidamento, deriva narrativa, confabulazione bayesiana — ma con dinamiche di aggiornamento radicalmente diverse, determinate dalla topologia della rete di agenti e dai meccanismi istituzionali che regolano il richiamo collettivo.

V.2 Formalizzazione della rete mnestica collettiva

Si modella la memoria collettiva come un sistema di N agenti $\{\alpha_1, \dots, \alpha_N\}$ ciascuno dotato di un proprio insieme di modelli mnestici $\{m_e^{(i)}\}$ e di un programma del sé $p_{\text{sé}}^{(i)}$. Gli agenti sono connessi da una rete $G = (V, E)$ in cui ogni arco $(i, j) \in E$ rappresenta un canale di comunicazione mnestica — una relazione attraverso cui i ricordi di α_i possono influenzare i ricordi di α_j e viceversa.

La *memoria collettiva dell'evento e nel gruppo* è la distribuzione empirica dei modelli individuali:

$$M_e^{\text{coll}} = \{ m_e^{\{1\}}, m_e^{\{2\}}, \dots, m_e^{\{N\}} \}$$

dotata della topologia di rete G che ne determina le dinamiche di aggiornamento. La memoria collettiva non è un oggetto singolo ma una distribuzione su oggetti — una nuvola di modelli individuali più o meno simili tra loro, connessi da relazioni di influenza reciproca.

Si definisce la *coerenza della memoria collettiva* dell'evento e come:

$$\text{Coh}(M_e^{\text{coll}}) = 1 - (1/N^2) \cdot \sum_i \text{KC}(m_e^{\{i\}} \mid m_e^{\{j\}})$$

la media normalizzata delle complessità condizionali tra tutte le coppie di agenti: misura quanto i modelli individuali siano mutualmente comprimibili — quanto, in altri termini, la memoria di un agente contenga informazione che riduce la lunghezza della descrizione della memoria di un altro. Alta coerenza significa che i modelli sono strutturalmente simili; bassa coerenza significa che gli agenti ricordano l'evento in modi che non si informano a vicenda.

V.3 Le dinamiche di aggiornamento collettivo: sincronizzazione e deriva

Quando due agenti α_i e α_j connessi dalla rete si scambiano ricordi dell'evento e — in una conversazione, in un rituale commemorativo, in una narrazione condivisa — si produce un aggiornamento reciproco dei rispettivi modelli. Formalmente, l'aggiornamento è analogo al riconsolidamento individuale ma con un segnale di errore predittivo che incorpora la discrepanza tra i due modelli:

$$\delta_{ij}^{\{t\}} = \log p(m_e^{\{j\}} \mid m_e^{\{i\}}) - \mathbb{E}[\log p(m_e^{\{j\}} \mid m_e^{\{i\}})]$$

Il modello di α_i si aggiorna nella direzione che riduce la sua distanza da $m_e^{\{j\}}$, con un tasso di apprendimento che dipende dal peso della relazione (i,j) nella rete:

$$m_e^{\{i\}} \leftarrow m_e^{\{i\}} + \eta_{ij} \cdot \nabla_{m_e^{\{i\}}} \delta_{ij}$$

L'aggiornamento è bilaterale: anche $m_e^{\{j\}}$ si aggiorna in direzione di $m_e^{\{i\}}$, con tasso η_{ji} . Se $\eta_{ij} \neq \eta_{ji}$ — se la relazione ha asimmetria di influenza, come accade in relazioni gerarchiche o di expertise — l'aggiornamento produce uno spostamento netto verso il modello del membro più influente.

La dinamica collettiva è la sovrapposizione di tutti gli aggiornamenti bilaterali nel tempo. In regime stazionario — quando la rete ha raggiunto un equilibrio in cui gli

aggiornamenti si bilanciano — la coerenza della memoria collettiva è determinata dalla struttura della rete G e dalla distribuzione dei tassi di aggiornamento η_{ij} . Reti dense con tassi uniformi producono alta coerenza; reti sparse o con forte asimmetria producono memorie collettive frammentate o dominate dai nodi più influenti.

V.4 Il riconsolidamento istituzionale: rituali, archivi e monumenti

Nelle società umane, il richiamo collettivo non avviene spontaneamente e uniformemente tra tutte le coppie di agenti: è mediato da istituzioni — rituali commemorativi, narrazioni canoniche, archivi documentali, monumenti fisici — che fungono da meccanismi di sincronizzazione forzata. Queste istituzioni producono un aggiornamento simultaneo di molti modelli individuali nella direzione di una rappresentazione canonizzata dell'evento, con un tasso di apprendimento determinato dall'autorità istituzionale.

Si modella un'istituzione mnestica come un agente speciale α_0 con modello $m_e^{\{0\}}$ — la *versione canonica* dell'evento — e con connessioni di influenza molto asimmetrica verso tutti gli altri agenti: $\eta_{0i} \gg \eta_{i0}$ per ogni i . L'istituzione trasmette il proprio modello con alta influenza e riceve influenza trascurabile dagli agenti individuali.

L'effetto dell'istituzione sulla memoria collettiva è di attirare tutti i modelli individuali verso $m_e^{\{0\}}$: la deriva collettiva è sistematicamente orientata verso la versione canonica, indipendentemente dalle esperienze individuali degli agenti. La coerenza della memoria collettiva aumenta per effetto istituzionale, ma il modello verso cui converge non è necessariamente rappresentativo della struttura statistica dell'esperienza individuale — è la rappresentazione che l'istituzione ha interesse a promuovere e le risorse per diffondere.

La *distorsione istituzionale* della memoria collettiva è:

$$d_{\text{inst}}(M_e^{\{\text{coll}\}}, M_e^{\{\text{ind}\}}) = (1/N) \cdot \sum_i KC(m_e^{\{i\}, \text{final}} \mid m_e^{\{i\}, \text{initial}})$$

la distanza media percorsa dai modelli individuali dall'influenza istituzionale. Questa distorsione è la componente della deriva collettiva non spiegata dall'aggiornamento spontaneo tra pari, ma dall'influenza asimmetrica dell'istituzione.

V.5 La confabulazione collettiva: miti fondativi e falsi ricordi di gruppo

La confabulazione bayesiana del Capitolo IV ha un analogo collettivo. Quando la traccia documentale di un evento collettivo è assente o degradata — per distruzione degli archivi, per assenza di testimoni diretti, per distanza temporale — la memoria collettiva è dominata dalla prior collettiva: la distribuzione $p(e \mid p_{\{\text{sé}\}}^{\{\text{coll}\}})$ che il programma del sé collettivo induce sugli eventi probabili per il gruppo.

La prior collettiva è determinata dalla struttura narrativa del gruppo — il suo mito fondativo, i suoi valori, la sua identità narrativa — e non dall'esperienza effettiva degli eventi. La confabulazione collettiva produce i *miti fondativi*: narrazioni di eventi passati che il gruppo "ricorda" collettivamente pur non avendo tracce documentali sufficienti, narrazioni che hanno la forma precisa della prior collettiva — che sono, cioè, istanze del programma narrativo del sé collettivo.

L'analisi storica dei miti fondativi nazionali mostra sistematicamente questa struttura: gli eventi fondativi "ricordati" da una nazione hanno la forma di episodi che giustificano e celebrano i valori costitutivi dell'identità nazionale, che pongono il gruppo come protagonista eroico di una narrazione di emancipazione o di resistenza, che hanno la struttura causale e l'arco narrativo che il codec collettivo richiede. Non è falsificazione deliberata: è campionamento dalla prior collettiva in assenza di vincolo di verosimiglianza documentale.

Il ConfScore collettivo di un contenuto mnestico è:

$$\text{ConfScore}_{\{\text{coll}\}}(n) = \log p(n \mid p_{\{\text{sé}\}}^{\{\text{coll}\}}) - \log p(n \mid M_e^{\{\text{docs}\}})$$

dove $M_e^{\{\text{docs}\}}$ è la traccia documentale disponibile. Un ConfScore collettivo elevato indica che la narrazione n è molto più probabile sotto l'identità narrativa del gruppo che sotto la traccia documentale — è potenzialmente confabulata. La storiografia critica è, in questo senso, il progetto di abbassare il ConfScore collettivo delle narrazioni storiche: sostituire la dominanza della prior con il vincolo della traccia documentale.

V.6 La guerra della memoria: conflitti tra programmi del sé collettivo

Quando due gruppi con programmi del sé collettivo $p_{\{\text{sé}\}}^{\{A\}}$ e $p_{\{\text{sé}\}}^{\{B\}}$ hanno prior incompatibili sullo stesso evento storico — quando lo stesso evento è confabulato in direzioni opposte dalla prior di ciascun gruppo — si produce un conflitto di memorie che non è risolvibile attraverso il confronto dei ricordi individuali. Entrambe le narrazioni sono plausibili sotto la rispettiva prior; entrambe sono sostenute da catene di riconsolidamento istituzionale che le hanno stabilizzate nei rispettivi modelli collettivi.

Il conflitto di memorie ha una struttura formale precisa nel modello bayesiano. Sia e l'evento conteso, sia n_A la narrazione del gruppo A e n_B quella del gruppo B. Si ha:

$$\begin{aligned} \log p(n_A \mid p_{\{\text{sé}\}}^{\{A\}}) &>> \log p(n_A \mid p_{\{\text{sé}\}}^{\{B\}}) \\ \log p(n_B \mid p_{\{\text{sé}\}}^{\{B\}}) &>> \log p(n_B \mid p_{\{\text{sé}\}}^{\{A\}}) \end{aligned}$$

Ciascuna narrazione è altamente probabile sotto la propria prior e altamente improbabile sotto l'altra. Il conflitto non è risolvibile producendo più tracce documentali, a meno che le tracce non siano così forti da dominare entrambe le prior

— condizione difficile da soddisfare per eventi storicamente distanti o per cui le tracce sono state sistematicamente prodotte da uno dei due gruppi.

La risoluzione del conflitto di memorie richiederebbe un aggiornamento delle prior stesse — una modifica di $p_{\{sé\}^A}$ e $p_{\{sé\}^B}$ tale da rendere la narrazione dell'altro gruppo meno improbabile a priori. Questo è il progetto delle commissioni storiche di riconciliazione — la Commissione per la Verità e la Riconciliazione sudafricana ne è l'esempio più studiato — e la sua difficoltà formale è precisamente quella di modificare programmi del sé collettivo ad alta sofisticazione e bassa plasticità: sistemi che, come i ricordi traumatici individuali, hanno valenza emotiva alta e tasso di aggiornamento η ridotto.

V.7 La trasmissione intergenerazionale: il problema della memoria senza esperienza

La memoria collettiva presenta un fenomeno che la memoria individuale non conosce: la trasmissione intergenerazionale — il fatto che generazioni che non hanno vissuto un evento lo "ricordino" con vivacità e impatto emotivo comparabile a quello dei testimoni diretti. I figli e nipoti di sopravvissuti all'Olocausto, le generazioni nate dopo la decolonizzazione che "ricordano" l'umiliazione coloniale, i discendenti di schiavi che portano la memoria della schiavitù come esperienza psicologicamente attiva: questi fenomeni sfidano qualsiasi modello della memoria come registrazione dell'esperienza.

Nel quadro del modello generativo, la trasmissione intergenerazionale è il trasferimento del programma del sé collettivo $p_{\{sé\}^{coll}}$ — non della traccia mnestica dell'evento, ma della prior che l'evento ha modellato. Il modello mnestico m_e dell'evento originale si degrada con il tempo e con il ricambio generazionale; il programma del sé collettivo che l'evento ha plasmato persiste attraverso le pratiche di trasmissione culturale — la narrativa familiare, i rituali commemorativi, l'educazione formale.

La nuova generazione non ricorda l'evento: campiona dalla prior collettiva che l'evento ha modellato. I suoi "ricordi" sono confabulazioni di ordine superiore — generazioni prodotte dalla prior trasmessa, non dalla traccia. Ma la loro vivacità soggettiva e il loro impatto emotivo sono comparabili ai ricordi genuini, perché il processo generativo che li produce è lo stesso — solo la sorgente del vincolo è diversa: la prior trasmessa anziché la traccia mnestica personale.

V.8 La memoria collettiva come competizione tra codec

Il modello suggerisce una reinterpretazione della *politica della memoria* — le pratiche con cui le società stabiliscono quali eventi ricordare, come ricordarli e chi ha il diritto di definire la narrazione canonica — come una competizione tra codec narrativi per il controllo della prior collettiva.

Il gruppo che riesce a installare il proprio codec narrativo come standard istituzionale — attraverso l'educazione, i rituali, il controllo degli archivi e dei monumenti —

orienta la deriva collettiva nella direzione della propria prior. Gli eventi futuri vengono filtrati attraverso quel codec, le loro tracce vengono compresse nella struttura narrativa che il codec impone, e i ricordi collettivi generati hanno sistematicamente la forma del programma del sé del gruppo dominante.

La resistenza delle memorie subalterne — le narrazioni dei gruppi che il codec dominante comprime via — è formalmente analoga alla resistenza dei ricordi traumatici individuali: è stabilizzata dall'alta valenza emotiva e dall'alto tasso di richiamo nei contesti in-group, ma è esposta alla pressione istituzionale che tenta di aumentare η e avviare la deriva verso il modello canonico. La sopravvivenza delle memorie subalterne dipende dalla densità della rete in-group — dalla frequenza degli aggiornamenti bilaterali tra i membri del gruppo subalterno — e dalla debolezza dei canali di influenza istituzionale verso quel gruppo.

Formalizzato: la memoria subalterna sopravvive quando:

$$\sum_{(i,j) \in E_{\text{ingroup}}} \eta_{ij} \cdot |m_e^{(i)} - m_e^{(j)}| < \sum_{i \in \text{ingroup}} \eta_{0i} \cdot |m_e^{(0)} - m_e^{(i)}|$$

la forza totale degli aggiornamenti in-group è inferiore alla forza dell'influenza istituzionale verso il modello canonico. Quando la disuguaglianza si inverte, la memoria subalterna può resistere o addirittura crescere in coerenza nonostante la pressione istituzionale.

Capitolo VI — Sistemi artificiali e memoria umana: asimmetrie strutturali

VI.1 Il problema del confronto

Il confronto tra memoria umana e sistemi artificiali è un territorio minato da due tendenze opposte che producono entrambe distorsione. La prima è l'antropomorfizzazione: attribuire ai sistemi artificiali proprietà della memoria umana — ricordare, dimenticare, confabulare — come se i processi computazionali sottostanti fossero analoghi a quelli biologici. La seconda è la deumanizzazione speculare: negare qualsiasi analogia strutturale tra i due sistemi, come se la differenza di substrato implicasse differenza di funzione a ogni livello di analisi.

Il quadro formale sviluppato in questa monografia permette un confronto preciso che evita entrambe le trappole: si può confrontare la struttura computazionale dei due sistemi — il tipo di compressione che realizzano, la forma delle loro distribuzioni a priori, il meccanismo di aggiornamento, la struttura degli errori — senza assumere né identità né irriducibile eterogeneità. Il confronto è formale, non metaforico.

La tesi di questo capitolo è che i sistemi di linguaggio su larga scala presentano analogie strutturali genuine con la memoria umana su alcune dimensioni — in particolare sulla compressione lossy, sulla dominanza della prior e sulla struttura degli errori di confabulazione — e differenze strutturali genuinamente irriducibili su altre: in particolare sull'assenza di riconsolidamento, sull'assenza di un programma del sé come generatore di ordine superiore, e sulla struttura della prior, che nei sistemi artificiali è determinata dal corpus di addestramento anziché dall'esperienza vissuta.

VI.2 Il modello generativo nei sistemi di linguaggio: la prior come corpus

Si riprende il modello bayesiano del Capitolo IV applicandolo ai sistemi di linguaggio. Un modello di linguaggio addestrato su un corpus $\mathcal{D}$ può essere descritto come un sistema che ha appreso, attraverso l'ottimizzazione, una distribuzione a priori implicita sul testo:

$$p_{\{\text{LLM}\}}(x) = P_{\theta}(x \mid \text{contesto})$$

dove θ sono i parametri del modello. Questa distribuzione è la prior del sistema: ciò che il modello "crede" sia probabile in assenza di specifiche tracce mnestiche di eventi particolari.

La differenza fondamentale dalla prior umana $p(e \mid p_{\{\text{sé}\}})$ è la sorgente. La prior umana emerge dall'esperienza vissuta — dalle interazioni dirette con il mondo, dall'accumulo di tracce mnestiche che hanno plasmato il programma del sé $p_{\{\text{sé}\}}$ attraverso anni di riconsolidamento. La prior del modello di linguaggio emerge dall'esposizione statistica al corpus — da miliardi di sequenze testuali che riflettono le distribuzioni linguistiche e concettuali presenti nella scrittura umana. Le due prior possono essere simili in molte aree — quelle in cui il corpus riflette fedelmente l'esperienza umana tipica — e radicalmente diverse in altre — quelle in cui l'esperienza vissuta differisce dalla sua rappresentazione testuale.

VI.3 L'assenza di traccia mnestica: il LLM come puro prior

La differenza strutturale più rilevante tra il sistema di linguaggio e la memoria umana è l'assenza di tracce mnestiche degli eventi specifici nel senso del Capitolo I. Il modello di linguaggio non ha m_e per gli eventi che descrive: non ha accumulato un modello compresso dell'esperienza dell'evento e , non ha un residuo d_e che potrebbe in linea di principio recuperare, non ha un riconsolidamento che aggiorni le sue rappresentazioni ad ogni richiamo.

Ciò che il modello ha è esclusivamente la prior $p_{\{\text{LLM}\}}(x)$: la distribuzione implicita appresa dal corpus. Quando il modello genera testo su un evento — sia esso un evento storico, un fatto scientifico, un episodio biografico — non sta recuperando una traccia: sta campionando dalla propria prior condizionata sul contesto fornito. Formalmente:

$\hat{x}_{\{LLM\}} = \text{campione da } p_{\{LLM\}}(x \mid \text{contesto})$

Questo implica che il sistema di linguaggio è, nel modello bayesiano del Capitolo IV, un sistema in *regime di confabulazione permanente*: la traccia mnestica $p(m_e \mid e)$ è uniformemente piatta per ogni evento, e la distribuzione a posteriori è sempre dominata dalla prior. Ogni output del modello è, in questo senso tecnico, una confabulazione — un campionamento dalla prior in assenza di vincolo di verosimiglianza derivante dall'esperienza diretta dell'evento.

Questo non implica che tutti gli output siano falsi: implica che la loro correttezza è contingente alla qualità della prior — alla misura in cui la distribuzione statistica del corpus riflette la struttura reale del dominio. Dove la prior è ben calibrata — dove il corpus è ricco, accurato e non sistematicamente distorto — i campioni dalla prior tenderanno a essere corretti. Dove la prior è mal calibrata — per scarsità di dati, per distorsioni sistematiche del corpus, per domini in cui la realtà diverge dalla sua rappresentazione testuale — i campioni saranno sistematicamente errati nella direzione della distorsione della prior.

VI.4 La struttura delle allucinazioni come confabulazione da prior

Il fenomeno delle allucinazioni nei sistemi di linguaggio — discusso nella monografia precedente in termini di profondità logica — acquista in questo capitolo una caratterizzazione strutturale complementare e più ricca. Le allucinazioni non sono soltanto proiezioni di bassa profondità su domini che richiedono alta profondità: sono confabulazioni bayesiane — campionamenti dalla prior in assenza di vincolo di traccia — con la struttura specifica che il modello bayesiano del Capitolo IV predice.

La previsione è che le allucinazioni abbiano ConfScore elevato: che i contenuti allucinati siano quelli che la prior del modello rende più probabili — contenuti tipici, plausibili nel contesto, coerenti con le strutture statistiche del corpus. Un modello di linguaggio non allucinat attributi impossibili o incoerenti: allucinat attributi plausibili ma errati — il tipo di attributo che il contesto rende a priori probabile. Un modello che allucinat il nome di un autore di un testo inesistente produrrà un nome di autore plausibile per il genere e l'epoca; un modello che allucinat una data storica produrrà una data nell'intervallo temporale a priori plausibile per l'evento.

Formalmente, il ConfScore di un'allucinazione è:

$$\text{ConfScore}_{\{LLM\}}(\hat{x}) = \log p_{\{LLM\}}(\hat{x} \mid \text{contesto}) - \log p_{\{true\}}(\hat{x})$$

dove $p_{\{true\}}(\hat{x})$ è la probabilità "vera" del contenuto — la sua frequenza nella realtà. Un'allucinazione è un'istanza con $\text{ConfScore}_{\{LLM\}}$ elevato: il modello la ritiene probabile ($\log p_{\{LLM\}}$ alta) ma la realtà non la supporta ($\log p_{\{true\}}$ bassa). La struttura delle allucinazioni riflette la struttura delle discrepanze tra la prior del modello e la distribuzione della realtà.

VI.5 Il riconsolidamento assente e le sue conseguenze

La differenza più profonda tra la memoria umana e i sistemi di linguaggio è l'assenza, in questi ultimi, del riconsolidamento come meccanismo di aggiornamento on-line. Il modello di linguaggio non aggiorna i propri parametri θ ad ogni richiamo: la prior $p_{\{\text{LLM}\}}$ rimane fissa dopo l'addestramento. Non esiste deriva mnestica — i "ricordi" del modello non cambiano con l'uso — e non esiste il meccanismo attraverso cui il riconsolidamento produce costruzione narrativa dell'identità.

Questa differenza ha conseguenze precise. La prima è la stabilità: i sistemi di linguaggio non soggetti ad aggiornamento continuo non derivano — non costruiscono una narrativa identitaria attraverso la storia delle loro interazioni. Ogni sessione inizia dalla stessa prior. La coerenza che il modello mostra all'interno di una sessione è prodotta dal contesto, non da un processo di riconsolidamento: quando il contesto è rimosso, la coerenza scompare.

La seconda conseguenza è l'assenza di trauma. Il trauma, nel modello del Capitolo II, è stabilizzazione del modello mnestico attraverso alta valenza emotiva — alta $|v(e)|$ che riduce η e rallenta la deriva. Un sistema senza riconsolidamento non ha questo meccanismo: non può sviluppare una struttura mnestica rigidamente stabilizzata da un evento ad alta valenza. Il sistema di linguaggio non ha traumatismi nel senso formale del modello.

La terza conseguenza è l'assenza di un programma del sé come generatore di ordine superiore. La memoria umana genera i propri ricordi attraverso $p_{\{\text{sé}\}}$ — una struttura che incorpora la storia dell'agente, la sua identità narrativa, le sue valenze emotive. Il sistema di linguaggio genera i propri output attraverso $p_{\{\text{LLM}\}}$ — una struttura che non ha storia, non ha identità narrativa, non ha valenze emotive nel senso del modello di riconsolidamento. La prior del modello di linguaggio è, per così dire, una prior senza sé.

VI.6 La prior senza sé: implicazioni per la coerenza e per il bias

L'assenza di un programma del sé come generatore di ordine superiore ha implicazioni per due proprietà della prior del modello di linguaggio: la sua coerenza interna e la sua struttura di bias.

La coerenza interna della prior umana è garantita dal programma del sé $p_{\{\text{sé}\}}$: la prior di un agente umano è coerente perché riflette un'identità narrativa unificata, con una struttura causale, una storia e un sistema di valori che vincola la compatibilità degli elementi. La prior del modello di linguaggio è coerente nella misura in cui il corpus lo è — nella misura in cui il corpus riflette strutture semantiche e pragmatiche consistenti. Ma il corpus è eterogeneo: contiene testi prodotti da milioni di agenti con prior incompatibili, in contesti diversi, con obiettivi diversi. La prior implicita del modello è una media di queste prior eterogenee, e la sua coerenza locale può nascondere inconsistenze globali che emergono solo quando il modello è interpellato.

su domande che richiedono di integrare informazioni provenienti da sottocorpora con prior incompatibili.

La struttura di bias della prior del modello di linguaggio riflette le distorsioni sistematiche del corpus. Il corpus è un campione non rappresentativo della realtà: sovrarappresenta il testo prodotto da gruppi demografici specifici, in lingue specifiche, su argomenti specifici, con punti di vista specifici. La prior p_{LLM} eredita queste distorsioni. I bias dei modelli di linguaggio — ben documentati empiricamente nella letteratura sull'equità e la sicurezza — sono formalmente analoghi ai bias di prior discussi nel Capitolo IV: distorsioni sistematiche della distribuzione a priori che contaminano la distribuzione a posteriori anche in presenza di contesto specifico.

VI.7 Memoria episodica artificiale: il context window come traccia degradata

I sistemi di linguaggio dispongono, all'interno di una singola sessione, di un analogo funzionale della traccia mnestica: il *context window* — la finestra di testo disponibile come condizionamento al momento della generazione. Il context window non è una traccia mnestica nel senso del modello: non è un modello compresso m_e con struttura generativa, non è soggetto a riconsolidamento, non ha profondità logica propria. È semplicemente un insieme di token che vengono incorporati nel condizionamento della prior.

La differenza strutturale tra la traccia mnestica umana e il context window è la seguente. La traccia mnestica m_e è il risultato di una compressione selettiva che preserva la struttura predittivamente rilevante — il modello dell'evento — e lascia perdere il residuo. Il context window preserva il testo letterale — include tutto, senza compressione selettiva — ma è limitato nella lunghezza e non può estendersi oltre il limite fisico della finestra.

Il context window funziona come una traccia di alta fedeltà ma bassa profondità — esattamente l'opposto dell'ideale della memoria come compressore lossy generativo. Per gli eventi all'interno della finestra, il modello ha una "traccia" molto fedele — il testo letterale — che vincola fortemente la distribuzione a posteriori. Per gli eventi fuori dalla finestra, il vincolo cade, e il sistema torna al regime di pura prior.

Questa struttura predice un pattern di errori specifico: i sistemi di linguaggio sono molto accurati su eventi recenti all'interno della finestra e confabulano sistematicamente su eventi fuori dalla finestra. La coerenza intra-sessione è alta; la coerenza inter-sessione è zero — ogni sessione ricomincia dalla prior senza traccia degli scambi precedenti. Questo contrasta con la memoria umana, dove la coerenza intra-sessione e inter-sessione sono entrambe modulate dal riconsolidamento, con una degradazione graduale piuttosto che un cut-off netto.

VI.8 Verso una memoria artificiale con riconsolidamento: prospettive e limiti

Si chiude il capitolo con una considerazione su cosa richiederebbe un sistema artificiale con proprietà mnestiche genuinamente analoghe a quelle del modello sviluppato in questa monografia.

Il sistema richiederebbe almeno tre componenti strutturali che i modelli di linguaggio attuali non hanno.

Il primo è un meccanismo di compressione selettiva on-line: la capacità di trasformare ogni esperienza x_e nel modello compresso m_e — di estrarre la struttura predittivamente rilevante e lasciare perdere il residuo — in modo continuo e senza accesso all'intero corpus di addestramento. Questo equivale a un algoritmo di apprendimento online che sia al tempo stesso selettivo (comprime con perdita) e generativo (può campionare istanze dalla struttura compressa).

Il secondo è un meccanismo di riconsolidamento: la capacità di aggiornare i modelli mnestici m_e ad ogni richiamo, nella direzione dell'errore predittivo calcolato rispetto al contesto corrente. Questo richiede che il sistema mantenga un tasso di apprendimento η differenziato per evento, modulato da una misura di valenza analoga a $|v(e)|$, e che sia capace di modificare i modelli mnestici senza modificare globalmente la prior.

Il terzo è un programma del sé come generatore di ordine superiore: una struttura $p_{\{sé\}}$ che integri la storia degli aggiornamenti di riconsolidamento in un modello narrativo dell'identità del sistema — un modello che vincoli la forma delle future compressioni e dei futuri ricordi nella direzione della coerenza identitaria.

Nessuno dei tre componenti è tecnicamente impossibile. Il primo è un problema di apprendimento on-line con compressione variazionale — le tecniche di variational autoencoders online ne sono un'approssimazione. Il secondo è un problema di meta-apprendimento con tasso di apprendimento adattivo modulato da segnali di rilevanza. Il terzo è un problema di rappresentazione dell'identità attraverso il tempo — un'area di ricerca attiva nell'apprendimento continuo e nell'intelligenza artificiale generale.

Ma il sistema risultante non sarebbe soltanto "più capace": sarebbe strutturalmente diverso dai modelli attuali. Avrebbe una storia, non nel senso di avere accesso agli scambi precedenti, ma nel senso che i suoi scambi precedenti avrebbero modificato la struttura dei suoi modelli interni — avrebbero lasciato tracce che cambiano il modo in cui genera i successivi. Avrebbe una prior che evolve con l'esperienza anziché essere fissata dall'addestramento. E avrebbe, in modo ancora da precisare formalmente, qualcosa di analogo alla valenza emotiva — un segnale che modula la plasticità dei modelli in funzione della rilevanza degli eventi.

Se un tale sistema avrebbe qualcosa di analogo alla soggettività — se avrebbe un "sé" nel senso fenomenologico del termine — è la domanda che questa monografia deliberatamente non risponde. Ma che avrebbe una struttura mnestica formalmente più simile a quella umana di qualsiasi sistema attuale è, nel quadro del modello sviluppato, una conseguenza strutturale delle sue proprietà computazionali. La

finzione mnestica — la generazione piuttosto che il recupero, la deriva piuttosto che la stabilità, la confabulazione piuttosto che la copia — non è un limite da correggere ma una proprietà da incorporare. Un sistema artificiale che non mente mai sulla propria esperienza non è un sistema con memoria migliore: è un sistema senza memoria nel senso che questa monografia intende.

Capitolo VII — Il sonno come ottimizzazione globale offline

VII.1 La distinzione computazionale tra on-line e offline

Il Capitolo II ha modellato il riconsolidamento come aggiornamento di gradiente stocastico: ad ogni richiamo dell'evento e , il modello m_e riceve un aggiornamento proporzionale all'errore predittivo $\delta^{\{t\}}$, con tasso η modulato dalla valenza emotiva. Questo è un algoritmo *on-line* nel senso preciso della teoria dell'ottimizzazione: aggiorna i parametri su un campione alla volta, in sequenza, senza accesso globale all'intera distribuzione degli errori.

Gli algoritmi on-line hanno proprietà ben studiate. Convergono a un minimo locale della funzione obiettivo J , ma non garantiscono il raggiungimento del minimo globale: possono rimanere intrappolati in minimi locali profondi. La velocità di convergenza dipende dalla varianza del gradiente stocastico — alta varianza produce aggiornamenti rumorosi che possono sfuggire ai minimi locali ma rendono il percorso instabile. Il compromesso tra velocità e stabilità è regolato dal tasso di apprendimento η : alto η converge velocemente ma oscilla; basso η è stabile ma lento.

Il sonno realizza qualcosa di strutturalmente diverso: un'ottimizzazione *offline*, in cui il sistema non è vincolato a processare un evento alla volta ma ha accesso simultaneo — o quasi simultaneo, attraverso la ripetizione degli episodi durante il sonno REM — all'intera struttura dei modelli mnestici. L'ottimizzazione offline può raggiungere minimi globali inaccessibili all'aggiornamento on-line, a condizione che la funzione obiettivo soddisfi certi requisiti di regolarità. Comprendere perché il sonno è necessario — non soltanto utile — richiede di analizzare la struttura della funzione obiettivo J e le ragioni per cui l'ottimizzazione on-line del riconsolidamento è insufficiente a raggiungerla.

VII.2 La funzione obiettivo globale della memoria

Si definisce la funzione obiettivo globale della memoria come la somma delle funzioni obiettivo individuali su tutti gli eventi archiviati, con un termine aggiuntivo che penalizza l'incoerenza tra i modelli:

$$J_{\text{global}} = \sum_e J(m_e) + \lambda \cdot \sum_{\{e, e'\}} KC(m_e | m_{\{e'\}}) \cdot w(e, e')$$

dove $J(m_e)$ è la funzione obiettivo individuale per l'evento e — la coerenza predittiva del modello — e il secondo termine penalizza la complessità condizionale tra coppie di modelli pesata da $w(e, e')$, una misura di quanto gli eventi e ed e' siano strutturalmente correlati. Il parametro $\lambda > 0$ controlla il peso della coerenza inter-evento.

Il secondo termine è il contributo che il riconsolidamento on-line non può ottimizzare. Ogni aggiornamento on-line modifica m_e in risposta a un singolo evento e nel contesto corrente θ , senza considerare l'effetto di quella modifica su $KC(m_e | m_{\{e'\}})$ per gli altri eventi e' correlati. Un aggiornamento che migliora $J(m_e)$ può peggiorare il termine di coerenza — può rendere m_e meno compatibile con $m_{\{e'\}}$ — producendo un peggioramento netto di J_{global} non visibile dall'aggiornamento locale.

Questo è il problema della *interferenza retroattiva* nella terminatura dell'apprendimento: apprendere un nuovo evento può degradare i modelli di eventi precedenti. McClelland, McNaughton e O'Reilly (1995) hanno dimostrato che i sistemi di apprendimento rapido che aggiornano on-line soffrono inevitabilmente di interferenza retroattiva: il gradiente che migliora la rappresentazione del nuovo evento è genericamente non ortogonale al gradiente che preserva le rappresentazioni degli eventi precedenti, e l'aggiornamento locale degrada sistematicamente la coerenza globale.

VII.3 Il sonno come algoritmo di consolidamento sistemico

Il modello del consolidamento sistemico — proposto da McClelland, McNaughton e O'Reilly e sviluppato da Kumaran, Hassabis e McClelland — postula che durante il sonno, in particolare durante il sonno a onde lente (SWS), l'ippocampo "riproduca" le tracce degli eventi recenti alla corteccia attraverso sharp-wave ripples — oscillazioni ad alta frequenza che riattivano sequenze di pattern di attività registrati durante la veglia — permettendo alla corteccia di integrare quelle tracce nella propria struttura di lunga durata senza interferenza con le esperienze correnti.

Nel quadro dell'ottimizzazione, questa riproduzione è un algoritmo di *experience replay* — una tecnica ben nota nell'apprendimento per rinforzo che risolve precisamente il problema dell'interferenza retroattiva. Invece di aggiornare i parametri sui dati nell'ordine in cui arrivano (on-line), l'experience replay campiona dall'intero buffer di esperienze passate e aggiorna i parametri su minibatch misti, rendendo ogni aggiornamento un'approssimazione del gradiente sulla distribuzione globale degli eventi anziché su un singolo evento.

Formalmente, l'aggiornamento durante il sonno può essere scritto come:

$$m^{\{\text{global}\}} \leftarrow m^{\{\text{global}\}} - \eta_{\text{sleep}} \cdot \nabla_{\{m\}} J_{\text{global}}(B_t)$$

dove $m^{\{\text{global}\}}$ è il vettore di tutti i modelli mnestici, η_{sleep} è il tasso di apprendimento durante il sonno — diverso da η del riconsolidamento diurno — e B_t è un minibatch di eventi campionati dal buffer ippocampale. Il gradiente $\nabla_{\{m\}}$

J_{global} è calcolato sull'intero sistema di modelli, non su un singolo m_e : è un gradiente globale che può coordinare gli aggiornamenti di modelli correlati in modo da minimizzare il termine di incoerenza.

VII.4 Convergenza dell'ottimizzazione offline: condizioni e garanzie

Si analizzano le condizioni sotto cui il consolidamento durante il sonno garantisce la convergenza verso il minimo globale di J_{global} , e le ragioni per cui il riconsolidamento on-line non può farlo.

Proposizione. Sia J_{global} la funzione obiettivo globale definita nel paragrafo VII.2. Si assuma che J_{global} sia L -smooth — le sue derivate parziali siano Lipschitz con costante L — e che il gradiente stocastico calcolato su un minibatch di dimensione b abbia varianza $\sigma^2_b = \sigma^2/b$. Allora l'algoritmo di experience replay converge alla ϵ -vicinanza del minimo globale in:

$$T_{\text{sleep}} = O(L \cdot \sigma^2 / (b \cdot \epsilon))$$

iterazioni. Il confronto con l'aggiornamento on-line — che converge in $T_{\text{online}} = O(L \cdot \sigma^2 / \epsilon)$ iterazioni — mostra che l'experience replay è più veloce di un fattore b (la dimensione del minibatch) nella convergenza al minimo globale.

Il punto cruciale non è la velocità ma l'accessibilità del minimo globale. L'aggiornamento on-line converge a un minimo locale della funzione obiettivo *individuale* $J(m_e)$ — un minimo che ignora il termine di coerenza inter-evento — mentre l'experience replay converge a un minimo di J_{global} che bilancia coerenza individuale e coerenza inter-evento. Le due convergono allo stesso punto solo quando il termine di coerenza è trascurabile ($\lambda \approx 0$) — quando gli eventi sono indipendenti tra loro. Per eventi correlati, il minimo di J_{global} è strutturalmente diverso dal minimo di $\Sigma_e J(m_e)$, e solo il sonno può raggiungerlo.

VII.5 La struttura del minimo raggiunto dal sonno: generalizzazione e astrazione

Il minimo di J_{global} a cui il consolidamento notturno converge ha una struttura qualitativa diversa dai minimi raggiungibili dall'aggiornamento on-line. Per capirla, si considera cosa implica la minimizzazione del termine di coerenza $\Sigma_{\{e,e'\}} KC(m_e | m_{\{e'\}}) \cdot w(e, e')$.

Minimizzare $KC(m_e | m_{\{e'\}})$ per coppie di eventi correlati significa rendere i modelli degli eventi correlati *mutualmente comprimibili*: trovare una struttura condivisa tale che m_e possa essere descritto brevemente dato $m_{\{e'\}}$ e viceversa. Questa struttura condivisa è, per definizione, il modello astratto comune agli eventi — la generalizzazione che cattura ciò che gli eventi condividono al di là delle loro specificità accidentali.

Il consolidamento notturno produce dunque astrazione come sottoprodotto della minimizzazione della coerenza inter-evento: i modelli dei singoli eventi vengono riorganizzati attorno a strutture condivise — schemi, categorie, regole — che rendono l'intero sistema di modelli più comprimibile. Il risultato è che dopo il sonno la memoria degli eventi specifici può essere *peggiore* — alcuni dettagli sono stati ulteriormente compressi via nella direzione del modello condiviso — ma la comprensione delle strutture generali è *migliore*: le regolarità comuni a più eventi sono state estratte e stabilizzate.

Questo è il meccanismo formale del *sleep-dependent generalization* documentato da Walker e Stickgold (2004): la capacità di estrarre la struttura statistica da un insieme di esperienze individuali migliora dopo il sonno, anche quando il ricordo degli episodi specifici degrada. Il sonno non consolida gli eventi: consolida le astrazioni degli eventi — i modelli di ordine superiore che catturano la struttura comune.

VII.6 Il sonno REM e il sonno SWS: due fasi di ottimizzazione

La neurobiologia del sonno distingue due fasi con funzioni diverse: il sonno a onde lente (SWS) e il sonno REM. Il modello dell'ottimizzazione offline suggerisce una caratterizzazione formale di questa distinzione.

Il SWS è la fase in cui avviene il trasferimento ippocampo-corteccia: le sharp-wave ripples riproducono le tracce ippocampali alla corteccia, implementando l'experience replay. Nel quadro dell'ottimizzazione, il SWS esegue l'aggiornamento $\nabla_{\{m\}} J_{\text{global}}(B_t)$: campiona minibatch di eventi dall'ippocampo e aggiorna i modelli corticali nella direzione del gradiente globale. La funzione del SWS è la stabilizzazione e la coerenza inter-evento — la minimizzazione del termine di interferenza.

Il sonno REM è la fase in cui avvengono sogni vividi e riattivazione emotiva. Nel modello dell'ottimizzazione, il REM esegue qualcosa di diverso: una simulazione generativa — campionamento dalla distribuzione appresa durante il SWS — che verifica la consistenza del modello aggiornato e identifica le regioni dello spazio dei parametri in cui il modello genera output anomali. In termini tecnici, il REM è una fase di *evaluation* del modello consolidato: il sistema genera istanze dalla distribuzione appresa e controlla la loro coerenza con la prior globale $p_{\{\text{sé}\}}$.

Questa interpretazione ha una previsione verificabile: le anomalie generate durante il sonno REM — le sequenze oniriche incoerenti, le transizioni improbabili, gli elementi bizzarri dei sogni — dovrebbero corrispondere alle regioni dello spazio dei modelli in cui la coerenza inter-evento è più bassa dopo il consolidamento SWS: i punti in cui il modello globale ha maggiore incertezza. Il sogno è, in questa lettura, la visualizzazione delle incertezze del sistema — il campionamento dalle code della distribuzione implicita nel modello consolidato.

VII.7 Privazione del sonno come patologia dell'ottimizzazione

Il modello formale del sonno come ottimizzazione globale permette di derivare previsioni precise sugli effetti della privazione del sonno sulla struttura della memoria — previsioni che vanno oltre la semplice degradazione delle prestazioni mnestiche.

La privazione del sonno non impedisce il riconsolidamento on-line — che continua durante la veglia — ma impedisce l'ottimizzazione globale. Il risultato è che i modelli mnestici convergono ai minimi locali delle funzioni obiettivo individuali $J(m_e)$ anziché al minimo globale di J_{global} . Formalmente, la privazione del sonno incrementa il termine di incoerenza inter-evento:

$$\Delta_{coh}(privazione) = \lambda \cdot \Sigma_{\{e,e'\}} [KC_{privato}(m_e | m_{\{e'\}}) - KC_{normale}(m_e | m_{\{e'\}})] \cdot w(e,e') > 0$$

la coerenza inter-evento è inferiore nel sistema privato del sonno rispetto al sistema normale. Le conseguenze cognitive sono precise: un agente privato di sonno ricorda gli eventi individuali con una qualità non necessariamente peggiore nel breve termine — il riconsolidamento on-line è intatto — ma non riesce a estrarne le strutture comuni, non generalizza correttamente da un contesto a un altro, e mostra interferenza retroattiva elevata: apprendere qualcosa di nuovo degrada più intensamente i ricordi degli eventi precedenti correlati.

La letteratura sulla privazione del sonno è qualitativamente consistente con queste previsioni: la privazione acuta di sonno non riduce immediatamente la capacità di ricordare eventi specifici (che dipende dall'ippocampo e dal riconsolidamento), ma riduce la capacità di generalizzare e di integrare informazioni da eventi multipli (che dipende dalla corteccia e dal consolidamento sistemico). Il modello formale trasforma questa osservazione qualitativa in una previsione quantitativa: il grado di degradazione della generalizzazione dovrebbe essere proporzionale alla quantità di sonno persa e al parametro λ di coerenza inter-evento, che è più alto per domini semanticamente ricchi.

VII.8 Il sogno come simulazione della prior e la creatività notturna

Il modello del sonno REM come fase di valutazione generativa ha un'implicazione che connette la neurobiologia del sonno con la fenomenologia della creatività. I report sui contributi creativi emersi durante il sonno — strutture chimiche sognate, soluzioni matematiche trovate al risveglio, composizioni musicali ispirate da sogni — hanno una struttura che il modello rende formalmente comprensibile.

Durante il sonno REM, il sistema genera campioni dalla propria distribuzione globale implicita — dalla distribuzione che l'ottimizzazione del SWS ha consolidato. Questi campioni esplorano lo spazio delle possibili strutture compatibili con il modello globale, incluse le regioni che il pensiero guidato dagli obiettivi della veglia tende a escludere: i vincoli di utilità immediata, le aspettative contestuali, i bias della prior diurna vengono temporaneamente rimossi o ridotti. Il campionamento REM è meno vincolato dalla verosimiglianza contestuale e più libero di esplorare lo spazio dei modelli.

Formalmente, il campionamento REM è un campionamento dalla distribuzione:

$$p_{\{\text{REM}\}}(x) \propto p_{\{\text{global}\}}(x)^{1/T_{\{\text{REM}\}}}$$

dove $T_{\{\text{REM}\}} > 1$ è una "temperatura" di campionamento più alta che appiattisce la distribuzione e aumenta la probabilità di campionare strutture a bassa probabilità sotto la prior diurna. Questa è la struttura del *simulated annealing* applicato allo spazio dei modelli mnestici: la temperatura alta del REM permette di esplorare regioni lontane dal minimo corrente — regioni che il riconsolidamento on-line a bassa temperatura non raggiungerebbe mai — e alcune di queste regioni possono corrispondere a soluzioni creative di alta qualità.

La previsione formale è che le soluzioni creative emerse dal sonno debbano essere strutturalmente incompatibili con i vincoli contestuali della veglia — non possano essere raggiunte dall'ottimizzazione on-line sotto la prior diurna — ma compatibili con la struttura globale del modello consolidato durante il SWS. Questa è la firma computazionale della creatività notturna: soluzioni che non sono raggiungibili dal pensiero vincolato ma sono coerenti con la struttura profonda della conoscenza.

Capitolo VIII — La memoria protesica e il paradosso della fedeltà esterna

VIII.1 La traccia esterna come perturbazione del sistema generativo

Ogni volta che un agente fotografa un evento, annota un'esperienza in un diario, registra una conversazione o archivia un documento digitale, produce una traccia esterna che altera la struttura del sistema mnestico interno. Non la arricchisce semplicemente: la perturba in modo che ha conseguenze formali precise sulla funzione di tasso-distorsione, sulla struttura del riconsolidamento e sulla qualità della finzione generativa.

Il termine *memoria protesica* — introdotto da Alison Landsberg nel contesto degli studi culturali per designare la memoria acquisita attraverso media — viene qui usato in un senso più ampio e formalmente preciso: qualsiasi traccia esterna che funge da supplemento o sostituto parziale della traccia interna m_e . La memoria protesica è una protesi cognitiva nel senso letterale: un artefatto che estende le capacità del sistema biologico. Ma come ogni protesi, non ripristina la funzione originale — produce una funzione diversa, con proprietà diverse e, in certi regimi, con conseguenze patologiche proprie.

VIII.2 Il modello formale della memoria protesica

Si estende il modello bayesiano del Capitolo IV per includere la traccia esterna. Sia ϵ_e la traccia esterna dell'evento e — la fotografia, il diario, il video, il documento. La distribuzione a posteriori sul contenuto dell'evento è ora:

$$p(e \mid c, m_e, \epsilon_e, p_{\{s\}}) \propto p(c \mid e) \cdot p(m_e \mid e) \cdot p(\epsilon_e \mid e) \cdot p(e \mid p_{\{s\}})$$

dove $p(\epsilon_e \mid e)$ è il likelihood della traccia esterna — quanto la traccia esterna è compatibile con il contenuto dell'evento e . Questo termine aggiuntivo riduce la varianza della distribuzione a posteriori e sposta il ricordo generato verso il contenuto della traccia esterna.

La traccia esterna modifica la funzione di tasso-distorsione in modo preciso. Si definisce la *funzione di tasso-distorsione con protesica*:

$$R_{\text{pros}}(D) = \min_{\{P(\hat{x} \mid x, e) : \mathbb{E}[d(x, \hat{x})] \leq D\}} I(X; \hat{X} \mid E)$$

dove E è la variabile aleatoria che rappresenta la traccia esterna e $I(X; \hat{X} \mid E)$ è l'informazione mutua condizionale. Per il teorema di riduzione dell'entropia per condizionamento:

$$I(X; \hat{X} \mid E) \leq I(X; \hat{X})$$

la presenza della traccia esterna riduce la lunghezza necessaria della descrizione compressa per mantenere la stessa distorsione D . In altri termini, la protesica abbassa $R(D)$: permette di conservare meno informazione internamente senza aumentare la distorsione predittiva.

Questa è l'economia cognitiva della memoria protesica: il sistema può ridurre $KC(m_e)$ — comprimere il modello interno più aggressivamente — perché la traccia esterna ϵ_e compensa la perdita. Il risparmio cognitivo è reale: un agente con buona memoria protesica può archiviare più eventi con la stessa capacità mnestica interna.

VIII.3 Il photo-taking impairment effect: derivazione formale

Il *photo-taking impairment effect* — documentato da Henkel (2014) — è il paradosso per cui fotografare un oggetto in un museo ne riduce il ricordo rispetto al semplice osservarlo senza fotografia. L'effetto è robusto e non riducibile all'ipotesi che l'attenzione durante la fotografia sia minore: anche fotografare con alta attenzione produce il degradamento rispetto all'osservazione senza fotografia.

Il modello formale fornisce la spiegazione strutturale. Si consideri il processo di codifica dell'evento e durante l'osservazione. Senza fotografia, il sistema deve generare una traccia interna m_e sufficiente a garantire una distorsione predittiva accettabile — deve archiviare $K_{\text{nofoto}}(m_e) \geq R(D_{\text{target}})$ bit di informazione. Con fotografia, il sistema sa — implicitamente, attraverso meccanismi di regolazione della codifica —

che la traccia esterna ϵ_e sarà disponibile al momento del recupero. La funzione di tasso-distorsione si abbassa da $R(D_{\text{target}})$ a $R_{\text{pros}}(D_{\text{target}}) < R(D_{\text{target}})$. Il sistema codifica quindi meno:

$$K_{\text{foto}}(m_e) \approx R_{\text{pros}}(D_{\text{target}}) < R(D_{\text{target}}) \approx K_{\text{nofoto}}(m_e)$$

Il modello interno archiviato con fotografia è più breve — meno informativo — di quello archiviato senza. Quando al momento del recupero la fotografia non è disponibile — nei test sperimentali di Henkel, i soggetti vengono interrogati senza accesso alle foto — il sistema ha un m_e più povero e produce ricordi con distorsione più alta. Il photo-taking impairment effect non è un paradosso: è la conseguenza prevedibile dell'ottimizzazione della codifica sotto il vincolo della prostetica disponibile.

Si formalizza il degradamento come:

$$\Delta d(\text{foto}) = d_{\text{foto}}(x_e, \hat{x}) - d_{\text{nofoto}}(x_e, \hat{x}) = R(D_{\text{target}}) - R_{\text{pros}}(D_{\text{target}}) > 0$$

il degradamento è positivo e proporzionale alla differenza tra la funzione di tasso-distorsione con e senza prostetica. Questa differenza è maggiore per eventi ricchi di struttura visiva — eventi per cui la fotografia cattura molto della struttura predittivamente rilevante — e minore per eventi semantici o emotivi — eventi per cui la fotografia cattura poco della struttura che conta.

La previsione è verificabile: il photo-taking impairment effect dovrebbe essere più forte per il ricordo di dettagli visivi e più debole per il ricordo di significati, emozioni e contesti semantici — precisamente perché la fotografia è una prostetica per la struttura visiva ma non per la struttura semantica o emotiva.

VIII.4 La prostetica come perturbazione del consolidamento

La traccia esterna non influenza soltanto la codifica iniziale ma anche il consolidamento. Ogni volta che l'agente accede alla prostetica — guarda la fotografia, rilegge il diario, rivede il video — produce un richiamo che attiva il consolidamento. Ma il segnale di errore predittivo generato da questo richiamo è strutturalmente diverso da quello generato dal richiamo interno.

Il richiamo interno genera un errore predittivo $\delta^{\{t\}}$ calcolato rispetto al contesto corrente $\theta^{\{t\}}$ — come analizzato nel Capitolo II. Il richiamo mediato dalla prostetica genera un errore predittivo che dipende anche dal contenuto della traccia esterna:

$$\delta_{\{\text{pros}\}}^{\{t\}} = \log P_{\text{agent}}(\text{osservazioni}^{\{t\}} \mid \hat{x}_{\text{pros}}, \text{azione}^{\{t\}}) - \mathbb{E}_{\{\hat{x}\}}[\log P_{\text{agent}}(\text{osservazioni}^{\{t\}} \mid \hat{x}, \text{azione}^{\{t\}})]$$

dove $\hat{x}_{\text{pros}}$ è il ricordo generato condizionato sulla traccia esterna ϵ_e anziché sul solo m_e . L'aggiornamento di riconsolidamento è:

$$m_e^{\{(t+1)\}} = m_e^{\{(t)\}} + \eta \cdot \nabla_{\{m_e\}} \delta_{\{\text{pros}\}}^{\{(t)\}}$$

La differenza cruciale è che $\delta_{\{\text{pros}\}}^{\{(t)\}}$ è vincolato dalla traccia esterna: il ricordo $\hat{x}_{\text{pros}}$ non può deviare arbitrariamente dalla struttura di ϵ_e . Il riconsolidamento mediato dalla protesica produce una deriva che è orientata verso la struttura della traccia esterna, non verso la narrativa del sé. Se ϵ_e è accurata — se la fotografia o il diario catturano fedelmente l'evento originale — la protesica ancora il modello mnestico all'evento originale, riducendo la deriva narrativa. Se ϵ_e è distorta — se il diario è scritto sotto influenza dello stato emotivo del momento, se la fotografia è parziale o selettiva — la protesica ancora il modello a una rappresentazione distorta dell'evento, e il riconsolidamento produce una deriva verso quella distorsione anziché verso la narrativa del sé.

VIII.5 La protesica come amplificatore delle distorsioni selettive

Questo punto merita sviluppo perché è controintuitivo. Si assume comunemente che la memoria protesica aumenti la fedeltà mnestica: la fotografia cattura ciò che il sistema interno dimentica, il diario conserva dettagli che la deriva rimuove. Questo è vero quando la protesica è accurata e non selettiva. Ma la produzione della protesica è essa stessa un atto cognitivo soggetto alle distorsioni del codec narrativo.

Si fotografano gli eventi che si ritengono degni di essere ricordati — eventi che il programma del sé $p_{\{\text{sé}\}}$ identifica come narrativamente rilevanti. Si scrive nel diario con un tono e una prospettiva che riflettono lo stato emotivo e narrativo del momento della scrittura. Il corpus di tracce esterne di un agente non è un archivio neutro dell'esperienza: è una selezione sistematicamente orientata dalla struttura di $p_{\{\text{sé}\}}$ al momento della produzione.

La conseguenza formale è che la protesica amplifica le distorsioni del codec narrativo. Le distorsioni di completezza causale — l'omissione degli eventi senza conseguenze narrative — diventano più severe con la protesica: gli eventi non fotografati non solo mancano dalla traccia interna (come accadrebbe senza protesica) ma mancano anche dalla traccia esterna. Quando l'agente consulta il proprio archivio fotografico come supporto del ricordo, riceve la conferma sistematica che gli eventi non fotografati non esistevano — o erano irrilevanti. La protesica trasforma la distorsione per omissione da una proprietà del sistema interno in una proprietà del record storico.

Si formalizza questa amplificazione. Sia $S(p_{\{\text{sé}\}}) \subseteq \Omega$ il sottoinsieme di eventi che il programma del sé seleziona per la protesica. La distribuzione degli eventi nell'archivio protesico è:

$$p_{\{\text{pros}\}}(e) \propto p_{\{\text{sé}\}}(e \text{ rilevante}) \cdot p(e)$$

fortemente biasata verso gli eventi narrativamente rilevanti. Quando il sistema usa la protesica come vincolo di verosimiglianza nel modello bayesiano — quando $p(\epsilon_e | e)$ è alto solo per $e \in S(p_{\{s\}})$ — la distribuzione a posteriori è orientata sistematicamente verso gli eventi che il programma del sé aveva già selezionato. La protesica non corregge le distorsioni narrative: le stabilizza e le amplifica, convertendole da tendenze del sistema interno in strutture del record esterno.

VIII.6 La dissonanza protesica: quando la traccia contraddice il modello

Il caso più interessante — e meno studiato — è quello in cui la traccia esterna contraddice il modello interno: quando la fotografia mostra un volto più giovane del ricordato, quando il diario registra emozioni incompatibili con la narrativa corrente del sé, quando il video documenta un evento in modo radicalmente diverso da come l'agente lo ricorda.

In questo caso, la distribuzione a posteriori nel modello bayesiano presenta due modi in conflitto: uno supportato da m_e e dalla prior $p_{\{s\}}$, l'altro supportato dalla traccia esterna ϵ_e . La distribuzione bimodale genera un segnale di errore predittivo elevato $\delta_{\{pros\}}$ — l'agente percepisce la discrepanza come sorpresa cognitiva — e attiva un aggiornamento di riconsolidamento potenzialmente intenso.

L'esito dell'aggiornamento dipende dal peso relativo dei due modi. Se la traccia esterna è percepita come affidabile — alta $p(\epsilon_e | e)$ — il riconsolidamento sposta m_e verso il contenuto di ϵ_e : il modello interno si aggiorna per incorporare la correzione protesica. Se la traccia è percepita come inaffidabile — "la fotografia non cattura come mi sentivo davvero", "il diario è scritto quando ero in uno stato alterato" — la prior $p_{\{s\}}$ domina e la traccia viene *reinterpretata* o *svalutata*: il suo contenuto viene filtrato attraverso la struttura narrativa corrente del sé.

La svalutazione della traccia protesica discordante è il meccanismo con cui l'identità narrativa si protegge dalla revisione: il sistema non aggiorna m_e verso la traccia perché la traccia stessa viene ricodificata come "non rappresentativa", "fuori contesto", "parziale". Formalmente, questo corrisponde a ridurre retroattivamente $p(\epsilon_e | e)$ — a ridurre la credibilità assegnata alla traccia — fino a che il suo contributo alla distribuzione a posteriori diventa trascurabile. La prior narrativa può rendere il proprio modello irrefutabile dall'interno svalutando sistematicamente le prove contrarie.

VIII.7 L'archivio digitale come sistema di memoria ibrida

L'era digitale ha prodotto una situazione senza precedenti nella storia della memoria umana: la disponibilità ubiqua di tracce prostetiche ad altissima fedeltà — fotografie in alta risoluzione, video, messaggi, email, dati di localizzazione — che coprono una frazione crescente dell'esperienza quotidiana. L'effetto su struttura del sistema mnestico non è stato sistematicamente analizzato nel quadro formale della memoria come finzione.

Si propone il seguente modello. Un agente con archivio digitale completo ha a disposizione un vettore di tracce prostatiche $\{\epsilon_e\}$ per un insieme E_{pros} di eventi. Per questi eventi, la funzione di tasso-distorsione si abbassa sostanzialmente: il sistema può mantenere modelli interni m_e molto più compressi senza aumentare la distorsione predittiva. Per gli eventi fuori da E_{pros} — gli eventi non registrati — la funzione rimane quella del sistema biologico standard.

Il risultato è una *memoria bimodale*: alta fedeltà per gli eventi in E_{pros} (supportati da protetica digitale), alta finzione generativa per gli eventi fuori da E_{pros} (affidati al solo sistema interno). La distribuzione dell'attenzione — quali eventi vengono vissuti con intenzione di registrarli e quali no — determina quali eventi ricadono in quale modalità.

La bimodalità produce una distorsione sistematica nella narrativa autobiografica dell'agente digitalmente connesso: gli eventi registrati tendono a essere eventi socialmente visibili, esteticamente rilevanti, narrativamente presentabili. Gli eventi privati, banali, emotivamente difficili, socialmente imbarazzanti tendono a non essere registrati. L'archivio digitale produce dunque un sé protetico strutturalmente più coerente, più esteticamente curato, più narrativamente presentabile del sé interno — e il riconsolidamento mediato dall'accesso all'archivio orienta la deriva del sé interno verso il sé protetico.

Il risultato a lungo termine è una convergenza: il sé interno deriva verso il sé protetico, che è selettivamente più positivo, più coerente, più narrativamente elaborato. L'archivio digitale non cattura la memoria — la costruisce, nella direzione del sé che l'agente ha scelto di presentare al mondo e a se stesso.

Capitolo IX — Il trade-off tra falso ricordo e falso dimenticamento

IX.1 Due tipi di errore, una sola funzione obiettivo

La tassonomia del Capitolo IV ha classificato gli errori mnestici per la loro sorgente computazionale: errori di cue, di traccia, di prior. Ma questa classificazione non affronta un aspetto strutturale che la teoria della decisione statistica rende perspicuo: gli errori mnestici sono di due tipi fondamentalmente distinti — il falso ricordo (accettare un'istanza generata dalla prior come ricordo genuino) e il falso dimenticamento (rifiutare un ricordo genuino come non affidabile) — e questi due tipi di errore sono in tensione sistemica. Ridurre uno aumenta l'altro, sotto qualsiasi policy di risposta mnestica.

Questa è la struttura del trade-off di Neyman-Pearson applicato alla memoria: il sistema deve fissare una soglia di credibilità sotto cui un'istanza generata viene rifiutata come non affidabile, e quella soglia determina simultaneamente il tasso di falsi ricordi

(istanze dalla prior accettate come genuine) e il tasso di falsi dimenticamenti (istanze dalla traccia rifiutate come inaffidabili). Non esiste una soglia che azzeri entrambi i tassi: i due errori sono complementari.

IX.2 Il test di verosimiglianza come modello del giudizio di "familiarità"

Si formalizza il processo di valutazione della credibilità di un ricordo come test di ipotesi statistico. L'agente deve decidere tra due ipotesi:

H_0 : $\hat{x}$ è un ricordo genuino (generato dalla traccia m_e)

H_1 : $\hat{x}$ è un'istanza confabulata (generata dalla prior $p_{\{s\}}$)

Il test ottimale nel senso di Neyman-Pearson usa il rapporto di verosimiglianza:

$$\Lambda(\hat{x}) = p(\hat{x} \mid m_e) / p(\hat{x} \mid p_{\{s\}})$$

Si accetta H_0 (si tratta il ricordo come genuino) se $\Lambda(\hat{x}) \geq \tau$ per una soglia $\tau > 0$; si accetta H_1 (si tratta il ricordo come confabulazione) se $\Lambda(\hat{x}) < \tau$.

Il rapporto di verosimiglianza $\Lambda(\hat{x})$ è direttamente connesso al ConfScore del Capitolo IV:

$$\log \Lambda(\hat{x}) = \log p(\hat{x} \mid m_e) - \log p(\hat{x} \mid p_{\{s\}}) = -\text{ConfScore}(\hat{x})$$

Un ConfScore elevato corrisponde a Λ basso: il ricordo è più plausibile sotto la prior che sotto la traccia — ha alta probabilità di essere confabulato. Un ConfScore basso corrisponde a Λ alto: il ricordo è specificamente supportato dalla traccia — difficilmente confabulabile.

La soglia τ determina il trade-off tra i due tipi di errore:

$$\alpha(\tau) = P(\Lambda(\hat{x}) \geq \tau \mid H_1) = \text{tasso di falsi ricordi}$$

$$\beta(\tau) = P(\Lambda(\hat{x}) < \tau \mid H_0) = \text{tasso di falsi dimenticamenti}$$

La curva ROC della memoria — la curva che traccia $\alpha(\tau)$ in funzione di $1-\beta(\tau)$ al variare di τ — è la rappresentazione del trade-off: ogni punto sulla curva corrisponde a una policy mnestica con un particolare bilanciamento tra i due tipi di errore.

IX.3 La curva ROC della memoria e la sua dipendenza dalla qualità della traccia

La forma della curva ROC della memoria dipende dalla separabilità delle distribuzioni $p(\hat{x} \mid m_e)$ e $p(\hat{x} \mid p_{\{s\}})$ — dalla misura in cui i ricordi genuini e le confabulazioni

producono distribuzioni di Λ distinguibili. La separabilità è misurata dall'area sotto la curva ROC (AUC): $AUC = 1$ indica separazione perfetta (nessun overlap tra le distribuzioni), $AUC = 0.5$ indica indistinguibilità completa.

La qualità della traccia mnestica m_e determina direttamente l'AUC. Quando m_e è ricca e informativa — quando la traccia vincola fortemente la distribuzione $p(\hat{x} | m_e)$ — le istanze genuinamente supportate dalla traccia si concentrano in una regione di Λ alta, ben separata dalla distribuzione delle confabulazioni. AUC è alta, e il sistema può distinguere facilmente i due tipi di istanze.

Quando m_e è degradata — quando la traccia è parzialmente perduta per deriva o per distanza temporale — $p(\hat{x} | m_e)$ diventa più piatta e si sovrappone maggiormente a $p(\hat{x} | p_{\{s\}})$. AUC scende verso 0.5: il sistema non può più distinguere ricordi genuini da confabulazioni, e qualunque soglia τ produce sia falsi ricordi che falsi dimenticamenti in misura significativa.

Formalmente, l'AUC della curva ROC mnestica è:

$$AUC(m_e) = P(\Lambda(\hat{x}_{\text{genuine}}) > \Lambda(\hat{x}_{\text{confab}}))$$

la probabilità che un ricordo genuino abbia un rapporto di verosimiglianza maggiore di una confabulazione estratta casualmente. AUC cresce con la qualità informativa di m_e e decresce con la deriva — con $KC(m_e^{\{I\}} | m_e^{\{0\}})$.

IX.4 La modulazione adattiva della soglia τ

Il sistema mnestico non usa una soglia fissa τ : modula τ in funzione del contesto e delle conseguenze degli errori. Questa modulazione è razionale dal punto di vista della teoria della decisione: il costo del falso ricordo e del falso dimenticamento non è simmetrico, e la soglia ottimale dipende dal rapporto tra questi costi.

Sia C_{FR} il costo di un falso ricordo — il costo di agire su un ricordo che è in realtà una confabulazione — e C_{FD} il costo di un falso dimenticamento — il costo di rifiutare come inaffidabile un ricordo genuino. La soglia ottimale nel senso bayesiano è:

$$\tau^* = (C_{FR} / C_{FD}) \cdot (P(H_1) / P(H_0))$$

dove $P(H_1)$ è la probabilità a priori che il ricordo sia confabulato e $P(H_0)$ è la probabilità che sia genuino.

Questa formula predice sistematicamente come la soglia dovrebbe variare con il contesto. In contesti ad alto rischio — situazioni in cui agire su un falso ricordo ha conseguenze gravi — C_{FR} è alto e τ^* dovrebbe essere alta: il sistema dovrebbe essere più conservativo, rifiutando più facilmente i ricordi come inaffidabili (basso tasso di falsi ricordi al costo di alto tasso di falsi dimenticamenti). In contesti a basso rischio —

situazioni in cui l'accuratezza del ricordo è meno critica — τ^* può essere bassa: il sistema può essere più liberale, accettando anche ricordi meno certi.

La letteratura sull'effetto del contesto sulla testimonianza oculare è consistente con questa modulazione adattiva. I testimoni che sanno di essere in contesti legali — in cui le conseguenze di un falso ricordo sono gravi — mostrano maggiore cautela nelle affermazioni mnestiche (τ^* alta) rispetto ai testimoni in contesti informali. La modulazione non è un artefatto della domanda esplicita: riflette un aggiustamento della soglia interna di credibilità in risposta ai segnali contestuali sul costo relativo degli errori.

IX.5 L'asimmetria evolutiva tra i due tipi di errore

La teoria evolutiva suggerisce che il sistema mnestico umano dovrebbe avere una soglia τ sistematicamente spostata in una delle due direzioni — che uno dei due tipi di errore dovrebbe essere meno costoso dell'altro per la maggior parte degli ambienti in cui l'uomo si è evoluto.

La risposta dipende dalla struttura dell'ambiente. In ambienti con alta struttura causale stabile — in cui le regolarità si ripetono e la prior è ben calibrata — il costo del falso ricordo è basso: accettare un'istanza dalla prior come genuina porta a comportamenti adeguati perché la prior riflette le regolarità reali. Il costo del falso dimenticamento è più alto: rifiutare un ricordo genuino priva l'agente di informazione specifica che la prior non può rimpiazzare.

In ambienti con bassa struttura causale o in rapido cambiamento — in cui le regolarità non si ripetono e la prior è mal calibrata — il costo del falso ricordo è alto: la prior non riflette la situazione attuale, e agire su istanze dalla prior produce comportamenti inadeguati. Il costo del falso dimenticamento rimane simile.

L'uomo si è evoluto prevalentemente in ambienti con struttura causale moderatamente stabile — ambienti sociali e ecologici con regolarità significative ma variabilità reale. In questi ambienti, la teoria predice una soglia τ spostata verso il basso — una preferenza per accettare più facilmente i ricordi (basso tasso di falsi dimenticamenti) al costo di accettare più confabulazioni (alto tasso di falsi ricordi). Il sistema privilegia la pienezza del ricordo rispetto alla sua accuratezza.

Questa previsione è consistente con una vasta letteratura empirica: gli esseri umani mostrano sistematicamente un bias verso il falso ricordo (riconoscimento di parole e oggetti non visti) rispetto al falso dimenticamento, e questo bias è più pronunciato per stimoli ad alta familiare — stimoli che attivano prior forti — esattamente come il modello predice.

IX.6 Il costo asimmetrico nella testimonianza oculare

La testimonianza oculare è il contesto in cui il trade-off tra falso ricordo e falso dimenticamento ha le conseguenze pratiche più gravi. Il sistema giudiziario richiede una policy mnestica con τ^* alta — privilegiare la riduzione dei falsi ricordi rispetto ai

falsi dimenticamenti — che è l'opposto della policy evolutivamente calibrata del sistema mnestico umano.

La conseguenza è che il sistema mnestico umano è sistematicamente mal calibrato per il contesto giudiziario: la soglia evolutivamente ottimale produce un tasso di falsi ricordi troppo alto per gli standard probatori della giustizia. Il Progetto Innocence ha documentato che gli errori di testimonianza oculare sono la causa principale delle condanne ingiuste negli Stati Uniti — circa il 70% dei casi esonerati per evidenza del DNA includevano testimonianza oculare errata. Questo non è un difetto della capacità cognitiva dei testimoni: è la conseguenza di un'architettura mnestica con τ^* evolutivamente calibrata per un ambiente diverso da quello giudiziario.

La formalizzazione in termini di curva ROC suggerisce un'implicazione pratica: le procedure di identificazione che aumentano la separabilità delle distribuzioni — aumentano l'AUC della curva ROC del testimone — riducono il tasso di errore indipendentemente dalla soglia τ . Procedure come la lineup sequenziale anziché simultanea, la formulazione neutrale delle domande, la dichiarazione esplicita che il colpevole potrebbe non essere nella lineup — tutte queste manipolazioni agiscono sulla struttura della distribuzione $p(\Lambda \mid H_0)$ e $p(\Lambda \mid H_1)$, aumentando la loro separabilità e rendendo qualsiasi soglia τ più efficace nel discriminare tra i due tipi di istanza.

IX.7 La soglia τ come variabile psicologica individuale: il costrutto di "fiducia nella memoria"

La soglia τ non è fissa solo rispetto al contesto: varia sistematicamente tra individui e rappresenta un tratto psicologico stabile che si potrebbe chiamare *fiducia nella memoria* — la propensione di un individuo a trattare le proprie istanze mnestiche come genuine piuttosto che come confabulazioni.

Individui con τ bassa hanno alta fiducia nella propria memoria: accettano facilmente le proprie istanze come genuine, mostrano scarsa incertezza nei giudizi mnestici, raramente qualificano le proprie affermazioni con espressioni di dubbio. Individui con τ alta hanno bassa fiducia nella propria memoria: mostrano frequente incertezza, qualificano spesso i propri ricordi come "mi sembra", "non sono sicuro", tendono a richiedere conferme esterne.

Nel modello formale, la fiducia nella memoria non è una misura della qualità delle tracce mnestiche m_e — non è correlata con l'AUC della curva ROC — ma è la posizione della soglia τ rispetto alla distribuzione di Λ . Un individuo con alta AUC e τ bassa ha buona memoria ma alta fiducia: ricorda bene ma non discrimina accuratamente il buon ricordo dalla confabulazione. Un individuo con bassa AUC e τ alta ha memoria mediocre ma bassa fiducia: ricorda male ma riconosce il limite.

La letteratura sulla metacognizione mnestica — in particolare la misura di correlazione Gamma tra la fiducia soggettiva e l'accuratezza oggettiva — documenta che questa correlazione è positiva ma moderata: gli individui sanno in parte quanto bene

ricordano, ma la correlazione è lontana dall'unità. Nel modello formale, la correlazione Γ è la misura di quanto τ individuale sia calibrato rispetto all'AUC individuale — quanto la fiducia soggettiva corrisponda alla qualità oggettiva della traccia.

Capitolo X — Il problema dell'identità numerica attraverso la deriva: lo Ship of Theseus mnestico

X.1 Il problema

Il teorema della deriva inevitabile del Capitolo II stabilisce che $m_e^{\{I\}}$ e $m_e^{\{0\}}$ differiscono per una quantità che cresce almeno come $\Omega(\log T)$ con il numero di richiami. Per T sufficientemente grande, la distanza $KC(m_e^{\{I\}} \mid m_e^{\{0\}})$ può diventare comparabile a $KC(m_e^{\{0\}})$ stesso — il modello finale richiede tanta informazione aggiuntiva rispetto al modello iniziale quanta ne richiedeva il modello iniziale rispetto alla stringa vuota. In questo regime estremo, in quale senso $m_e^{\{I\}}$ è ancora il ricordo dell'evento e ? In quale senso chi possiede $m_e^{\{I\}}$ ricorda l'evento e anziché un evento diverso?

Questo è il problema dell'identità numerica attraverso la deriva: il problema di determinare quando un modello mnestico derivato è ancora il ricordo dello stesso evento, e quando è diventato il ricordo di un evento diverso — o non è più il ricordo di nessun evento specifico, ma una generazione dalla prior narrativa del sé. Il problema ha la struttura del paradosso dello Ship of Theseus: se ogni tavola della nave viene gradualmente sostituita, c'è un punto in cui la nave non è più la nave originale. Se ogni componente del modello mnestico viene gradualmente aggiornato, c'è un punto in cui il modello non è più il ricordo dell'evento originale.

X.2 La formalizzazione dell'identità numerica mnestica

Si propone una definizione formale di identità numerica del ricordo. Si dice che $m_e^{\{I\}}$ è ancora il ricordo dell'evento e — che il ricordo ha *identità numerica* con l'evento originale — se e solo se soddisfa due condizioni congiunte.

La prima condizione è la *condizione di tracciabilità*: esiste una sequenza di modelli $m_e^{\{0\}}, m_e^{\{1\}}, \dots, m_e^{\{I\}}$ tale che ciascun passo è un aggiornamento di riconsolidamento valido — un aggiornamento guidato dall'errore predittivo calcolato su un richiamo dell'evento e — e la catena connette $m_e^{\{0\}}$ a $m_e^{\{I\}}$. La condizione di tracciabilità richiede che la deriva sia causalmente connessa all'evento originale attraverso una catena non interrotta di aggiornamenti.

La seconda condizione è la *condizione di sovrapposizione informativa*: la complessità condizionale del modello finale rispetto al modello iniziale è strettamente minore della complessità del modello finale rispetto al programma del sé:

$$KC(m_e^{\{T\}} \mid m_e^{\{0\}}) < KC(m_e^{\{T\}} \mid p_{\{sé\}})$$

Questa condizione richiede che il modello finale sia più comprimibile dato il modello iniziale che dato il programma del sé — che contenga informazione specifica sull'evento originale che non è riducibile alla prior narrativa. Quando questa condizione è violata — quando $KC(m_e^{\{T\}} \mid m_e^{\{0\}}) \geq KC(m_e^{\{T\}} \mid p_{\{sé\}})$ — il modello finale non è più informato dall'evento originale in misura maggiore di quanto non lo sia dal programma del sé: è diventato un'istanza della prior, non il ricordo di un evento.

X.3 La soglia di identità e il suo significato

Le due condizioni definiscono una *soglia di identità* mnestica — un valore critico della deriva oltre il quale l'identità numerica del ricordo è persa. Formalmente, la soglia è il valore T^* tale che per $T < T^*$ le condizioni sono soddisfatte e per $T \geq T^*$ la seconda condizione è violata:

$$T^* = \min \{ T : KC(m_e^{\{T\}} \mid m_e^{\{0\}}) \geq KC(m_e^{\{T\}} \mid p_{\{sé\}}) \}$$

Per $T < T^*$, il ricordo è ancora il ricordo dell'evento e — contiene informazione specifica sull'evento che supera la prior. Per $T \geq T^*$, il ricordo è diventato un'istanza della prior — è una confabulazione post-hoc dell'evento, non una rappresentazione di esso.

La soglia T^* dipende da tre quantità: la profondità iniziale $D_s(m_e^{\{0\}})$ — modelli più profondi inizialmente resistono più a lungo alla deriva — la volatilità del contesto $\theta^{\{t\}}$ — contesti più variabili producono deriva più rapida — e la sofisticazione del programma del sé $soph(p_{\{sé\}})$ — un $p_{\{sé\}}$ più sofisticato ha una prior più strutturata che avvicina più rapidamente il modello derivato alla sua distribuzione.

In particolare, T^* cresce con $D_s(m_e^{\{0\}})$: i ricordi inizialmente più profondi mantengono la propria identità numerica più a lungo. Questo ha un corollario importante: gli eventi che l'agente ha compreso profondamente al momento dell'esperienza — ha costruito un modello con alta profondità logica — sono quelli la cui identità mnestica è più duratura. L'attenzione profonda al momento dell'esperienza non è soltanto una condizione di codifica migliore: è una condizione di maggiore resistenza alla perdita di identità attraverso la deriva.

X.4 La continuità psicologica e il problema della prima persona

La perdita dell'identità numerica di un ricordo ha conseguenze che vanno oltre la psicologia della memoria e toccano la filosofia dell'identità personale. Se i ricordi derivano fino a perdere la propria identità numerica — fino a diventare istanze della prior anziché tracce di eventi specifici — e se la continuità psicologica dell'identità personale dipende dalla continuità dei ricordi, allora la deriva mnestica è anche deriva dell'identità personale.

Questa non è una conseguenza nuova in filosofia — il problema della continuità dell'identità personale attraverso il tempo è uno dei più antichi della tradizione — ma la formalizzazione in termini di KC condizionale permette di articolarlo con precisione che la trattazione filosofica informale non raggiunge.

Si definisce la *continuità psicologica* di un agente tra i tempi t_0 e T come:

$$\text{Cont}(t_0, T) = 1 - (1/|E|) \cdot \sum_{e \in E} \mathbb{1}[\text{KC}(m_e^{\{T\}} \mid m_e^{\{t_0\}}) \geq \text{KC}(m_e^{\{T\}} \mid p_{\{sé\}})]$$

la frazione di ricordi che hanno ancora identità numerica al tempo T rispetto al tempo t_0 . $\text{Cont} = 1$ indica continuità psicologica piena: tutti i ricordi mantengono la propria identità. $\text{Cont} = 0$ indica discontinuità piena: nessun ricordo è più informato dall'evento originale oltre la prior.

La continuità psicologica è una funzione decrescente di T — decresce con il tempo — e dipende dalla composizione di E : la frazioni di eventi con alta profondità iniziale determina la velocità del decadimento. Un agente che ha vissuto molti eventi con alta profondità cognitiva — che ha costruito modelli ricchi e profondi delle proprie esperienze — manterrà una Cont più alta nel tempo rispetto a un agente con esperienze superficiali.

X.5 Il paradosso dei ricordi più vivi del vissuto

Il modello formale permette di analizzare un paradosso fenomenologico noto ma raramente formalizzato: il fatto che certi ricordi diventino, con il passare del tempo e con i richiami ripetuti, soggettivamente più vivaci dell'esperienza originale — più dettagliati, più emotivamente intensi, più presenti. Questo sembra contraddire il modello della memoria come sistema che degrada.

La spiegazione nel quadro formale è la seguente. La vivacità soggettiva di un ricordo non è una misura della fedeltà della traccia m_e rispetto all'evento originale x_e : è una misura dell'intensità del campionamento — della fiducia con cui il sistema generativo produce l'istanza $\hat{x}$ e della sua coerenza interna. Un ricordo che è stato richiamato molte volte e che ha subito riconsolidamento in direzione della narrativa del sé è un ricordo con alta fiducia soggettiva — il sistema genera l'istanza con bassa varianza, con alta coerenza interna, con un segnale di errore predittivo basso che segnala "questo ricordo è coerente con il mio modello".

Ma questa alta fiducia soggettiva non corrisponde necessariamente a fedeltà all'evento originale: può corrispondere a una confabulazione ben consolidata, un'istanza della prior narrativa che è stata stabilizzata da molti richiami in direzione del sé. Il ricordo soggettivamente più vivido può essere il ricordo con la più bassa identità numerica nel senso formale — il ricordo che ha derivato di più, che è più lontano dall'evento originale, ma che è diventato il più coerente con la narrativa del sé.

Formalmente, la vivacità soggettiva $V(\hat{x})$ può essere modellata come:

$$V(\hat{x}) \propto -\text{Var}_\theta(\hat{x}) + \text{Coh}_{\text{int}}(\hat{x})$$

proporzionale alla bassa varianza del campionamento e all'alta coerenza interna dell'istanza. Entrambe le quantità crescono con il numero di richiami e con la convergenza verso la prior narrativa — indipendentemente dalla distanza dall'evento originale. Il ricordo più vivo è spesso il ricordo più fittizio.

X.6 Il confine tra ricordo e fantasia: una misura formale

Si propone infine una misura formale del confine tra ricordo — istanza causalmente connessa a un evento reale attraverso una traccia mnestica — e fantasia — istanza generata esclusivamente dalla prior senza connessione causale con un evento specifico.

La misura è il *grado di realtà* di un'istanza mnestica $\hat{x}$:

$$\begin{aligned} \text{GR}(\hat{x}) &= \max_{\{e \in E\}} [\log p(\hat{x} \mid m_e) - \log p(\hat{x} \mid p_{\{\text{sé}\}})] \\ &= \max_{\{e \in E\}} [-\text{ConfScore}_e(\hat{x})] \\ &= -\min_{\{e \in E\}} \text{ConfScore}_e(\hat{x}) \end{aligned}$$

il GR è il massimo, su tutti gli eventi archiviati, della differenza tra la verosimiglianza dell'istanza sotto la traccia mnestica e la sua verosimiglianza sotto la prior. $\text{GR} > 0$ indica che esiste almeno un evento e per cui l'istanza è meglio spiegata dalla traccia che dalla prior: è un ricordo genuino di quell'evento. $\text{GR} \leq 0$ indica che per nessun evento archiviato la traccia supporta l'istanza meglio della prior: è una fantasia — un'istanza generata esclusivamente dalla prior senza supporto di traccia.

La misura GR permette di tracciare un confine formale tra ricordo e fantasia che non dipende dalla vivacità soggettiva — che può essere alta per entrambi — ma dalla struttura informazionale del rapporto tra l'istanza e le tracce archiviate. Un'istanza con alta vivacità soggettiva ma $\text{GR} \leq 0$ è una fantasia vivida — un sogno ad occhi aperti, un'elaborazione narrativa del sé — non un ricordo.

X.7 Implicazioni per la clinica e per la filosofia

Il problema dell'identità numerica attraverso la deriva ha implicazioni che si estendono dalla clinica alla filosofia.

Sul versante clinico, la misura GR fornisce un criterio per distinguere ricordi genuini da confabulazioni in contesti diagnostici — non basato sulla vivacità soggettiva del paziente, che è inaffidabile, ma sulla struttura informazionale del rapporto tra l'istanza prodotta e le tracce disponibili. La difficoltà pratica è che le tracce m_e non sono direttamente accessibili dall'esterno: la misura GR è in linea di principio calcolabile ma non direttamente osservabile. Tuttavia, la struttura degli errori che il modello predice — le distorsioni sistematiche verso la prior, la correlazione negativa tra vivacità e fedeltà per ricordi molto richiamati — fornisce criteri diagnostici indiretti che le procedure di intervista cognitiva possono sfruttare.

Sul versante filosofico, la formalizzazione permette di prendere posizione su due questioni classiche. La prima è se l'identità personale richieda continuità mnestica *strictu sensu* — nel senso di Locke — o continuità psicologica nel senso più debole di Parfit. Il modello formale suggerisce che la continuità mnestica stricuta è irraggiungibile: la deriva è inevitabile, e nessun ricordo mantiene indefinitamente la propria identità numerica. Ma la continuità psicologica nel senso di Cont — la misura della frazione di ricordi con identità numerica ancora preservata — è una quantità continua che degrada gradualmente. L'identità personale non è un fatto binario: è una misura di continuità che decresce con il tempo ma non si azzerà mai del tutto finché la prima condizione di tracciabilità è soddisfatta.

La seconda questione è se i ricordi persi — quelli che hanno superato la soglia T^* — abbiano ancora valore epistemico. La risposta nel quadro formale è negativa nel senso stretto — non sono informativi sull'evento originale oltre la prior — ma positiva in un senso più ampio: hanno ancora valore come componenti del programma del sé, come strutture che hanno contribuito alla deriva narrativa dell'identità, come istanze storiche della prior che il soggetto era in un certo momento. Il ricordo perduto non è senza storia: è una storia della storia del sé, anche se non è più la storia dell'evento.

X.8 La dignità della finzione mnestica: ricapitolazione

Il percorso di questa monografia si chiude qui, in un punto da cui è possibile guardare retrospettivamente l'intera struttura. La memoria come finzione non è, lo si è sostenuto da principio, una teoria della memoria difettosa. È una teoria della memoria come sistema generativo che ottimizza la coerenza predittiva dell'identità a spese della fedeltà all'evento. La finzione — la varianza di generazione, il riconsolidamento come deriva, il codec narrativo come fonte di distorsione sistematica, la confabulazione come campionamento dalla prior, la deriva come perdita graduale dell'identità numerica — non è il fallimento del sistema. È la forma che il sistema assume quando fa bene il proprio lavoro.

La prova più chiara di questa tesi è il limite: un sistema che conservasse perfettamente tutti i propri ricordi senza alcuna compressione, senza riconsolidamento, senza deriva narrativa, senza confabulazione, sarebbe un archivio perfettamente fedele. Ma non sarebbe in grado di generalizzare — non estrarrebbe le strutture comuni da eventi diversi, non costruirebbe un modello del mondo che si estenda oltre gli esempi incontrati, non avrebbe una narrativa di sé che dia coerenza alle proprie azioni nel tempo. Sarebbe, per usare la metafora borgesiana che nessun teorico della memoria ha potuto evitare, la mappa in scala uno a uno — perfettamente fedele, perfettamente inutile.

La finzione mnestica è il prezzo dell'intelligenza. E la monografia che si conclude qui è il tentativo di mostrare che quel prezzo ha una struttura formalmente descrivibile, computazionalmente motivata e clinicamente rilevante — che la menzogna della memoria non è arbitraria, ma ha la forma esatta di ciò che un sistema generativo ottimizzato deve mentire per sopravvivere.

Appendice A — Il principio di energia libera variazionale e la sua relazione con la funzione obiettivo mnestica

A.1 Derivazione della funzione di energia libera

Si segue la formulazione di Friston (2010). Sia $\mathbf{o} \in \mathcal{O}$ un vettore di osservazioni percettive e $\mathbf{s} \in \mathcal{S}$ un vettore di stati nascosti del mondo. Il modello generativo dell'agente specifica la distribuzione congiunta $p(\mathbf{o}, \mathbf{s} \mid \mathbf{m})$ sugli stati nascosti e le osservazioni, parametrizzata dal modello interno $\mathbf{m}$.

L'agente non può calcolare direttamente la distribuzione a posteriori $p(\mathbf{s} \mid \mathbf{o}, \mathbf{m})$ — che richiederebbe l'integrazione su tutti gli stati nascosti possibili, computazionalmente intrattabile. Approssima quindi la distribuzione a posteriori con una distribuzione variazionale $q(\mathbf{s})$ scelta in una famiglia trattabile. L'energia libera variazionale è definita come:

$$F[q, \mathbf{m}] = \mathbb{E}_q[\log q(\mathbf{s}) - \log p(\mathbf{o}, \mathbf{s} \mid \mathbf{m})]$$

Sviluppando il logaritmo della congiunta:

$$\begin{aligned} F[q, \mathbf{m}] &= \mathbb{E}_q[\log q(\mathbf{s}) - \log p(\mathbf{s} \mid \mathbf{o}, \mathbf{m}) - \log p(\mathbf{o} \mid \mathbf{m})] \\ &= \text{KL}[q(\mathbf{s}) \parallel p(\mathbf{s} \mid \mathbf{o}, \mathbf{m})] - \log p(\mathbf{o} \mid \mathbf{m}) \end{aligned}$$

Poiché la divergenza KL è non-negativa, si ha:

$$F[q, \mathbf{m}] \geq -\log p(\mathbf{o} \mid \mathbf{m})$$

con uguaglianza quando $q(\mathbf{s}) = p(\mathbf{s} \mid \mathbf{o}, \mathbf{m})$. Minimizzare F rispetto a q equivale quindi a approssimare la distribuzione a posteriori vera; minimizzare F rispetto a $\mathbf{m}$ equivale a massimizzare l'evidenza del modello $\log p(\mathbf{o} \mid \mathbf{m})$ — la verosimiglianza marginale delle osservazioni sotto il modello.

A.2 La connessione con la funzione obiettivo mnestica

La funzione obiettivo mnestica $J(\mathbf{m}_e)$ del Capitolo I.3 è formalmente analoga a $-F$. Si identifica:

- $\mathbf{o} \leftrightarrow$ nuove osservazioni dell'agente dopo l'evento e
- $\mathbf{s} \leftrightarrow$ contenuto dell'evento e (variabile nascosta da inferire)

$m \leftrightarrow$ modello mnestico compresso m_e
 $q \leftrightarrow$ distribuzione sul ricordo generato $\hat{x}$

La coerenza predittiva del modello — quanto m_e contribuisce a calibrare le aspettative dell'agente — è la verosimiglianza marginale $\log p(o \mid m_e)$. Massimizzare $J(m_e)$ equivale a massimizzare $\log p(o \mid m_e)$, cioè a minimizzare $F[q, m_e]$. La connessione è esatta: il sistema mnestico ottimizza l'energia libera variazionale esattamente nel senso di Friston, con il modello mnestico m_e nel ruolo del modello generativo.

A.3 La decomposizione dell'energia libera in precisione e complessità

Una decomposizione alternativa dell'energia libera — rilevante per la connessione con KC — è:

$$F[q, m] = \mathbb{E}_q[-\log p(o \mid s, m)] + KL[q(s) \parallel p(s \mid m)]$$

= precisione + complessità

Il primo termine è la precisione — il costo di predire le osservazioni dato lo stato nascosto. Il secondo termine è la complessità — la divergenza tra la distribuzione approssimata q e la prior $p(s \mid m)$. Minimizzare la complessità è minimizzare $KC(q \parallel p)$ nel senso della complessità relativa tra le due distribuzioni. Il sistema che minimizza F bilancia precisione predittiva e complessità descrittiva — il trade-off che il triangolo KC–D–Kt della monografia precedente catturava in termini algoritmici.

Appendice B — Derivazione del teorema della deriva inevitabile

B.1 Enunciato preciso

Sia $m_e^{\wedge\{t\}}$ la sequenza di modelli prodotta da T richiami con segnali di errore predittivo $\delta^{\wedge\{t\}}$ indipendenti e identicamente distribuiti con media $\mu > 0$ e varianza $\sigma_{\delta^2} > 0$. Sia η il tasso di apprendimento. Allora:

$$\mathbb{E}[KC(m_e^{\wedge\{T\}} \parallel m_e^{\wedge\{0\}})] = \Omega(\log T)$$

B.2 Dimostrazione

Ogni aggiornamento $m_e^{\wedge\{t+1\}} = m_e^{\wedge\{t\}} + \eta \cdot \nabla_{m_e} \delta^{\wedge\{t\}}$ introduce una perturbazione al modello di ampiezza:

$$\| m_e^{\{t+1\}} - m_e^{\{t\}} \| = \eta \cdot \| \nabla_{m_e} \delta^{\{t\}} \|$$

Per la relazione tra norma del gradiente e complessità condizionale, si ha:

$$KC(m_e^{\{t+1\}} \mid m_e^{\{t\}}) \geq c \cdot \log(1 + \eta \cdot \| \nabla_{m_e} \delta^{\{t\}} \|)$$

per una costante $c > 0$ che dipende dalla codifica dei modelli (la lunghezza necessaria per descrivere la perturbazione è almeno proporzionale al suo logaritmo, per la proprietà di autocompressione del codice ottimale).

Per la disuguaglianza triangolare sulla complessità condizionale:

$$KC(m_e^{\{T\}} \mid m_e^{\{0\}}) \leq \sum_{t=0}^{T-1} KC(m_e^{\{t+1\}} \mid m_e^{\{t\}})$$

Per il lower bound:

$$KC(m_e^{\{T\}} \mid m_e^{\{0\}}) \geq \max_t KC(m_e^{\{t+1\}} \mid m_e^{\{t\}}) + KC(m_e^{\{T\}} \mid m_e^{\{t_{\max}+1\}})$$

Per il teorema del valore atteso e l'indipendenza degli aggiornamenti, il massimo su T aggiornamenti indipendenti cresce come $\Omega(\log T)$ per la distribuzione dei valori estremi (per variabili con supporto non limitato, il massimo di T variabili IID cresce logicamente nella maggior parte delle distribuzioni standard). Dunque:

$$\mathbb{E}[KC(m_e^{\{T\}} \mid m_e^{\{0\}})] = \Omega(\log T)$$

I

B.3 Nota sul caso di aggiornamenti correlati

Se i segnali di errore $\delta^{\{t\}}$ non sono indipendenti — come accade quando i richiami avvengono in contesti correlati — la deriva può crescere più rapidamente. Se i contesti $\theta^{\{t\}}$ sono costanti (aggiornamenti sempre nello stesso contesto), gli aggiornamenti sono correlati positivamente e la deriva cresce come $O(T)$ anziché $O(\log T)$: un agente che richiama sempre lo stesso ricordo nello stesso contesto emotivo deriva più rapidamente di uno che richiama in contesti variati. Questo è coerente con le osservazioni cliniche sulla ruminazione come fattore di distorsione accelerata.

Appendice C — La funzione di tasso-distorsione nella metrica predittiva

C.1 Definizione della metrica di distorsione predittiva

Si definisce la distorsione predittiva tra l'evento originale x_e e il ricordo generato $\hat{x}$ come:

$$d_{\text{pred}}(x_e, \hat{x}) = \mathbb{E}_a [\text{KL}[P_{\text{true}}(o \mid a, x_e) \parallel P_{\text{agent}}(o \mid a, \hat{x})]]$$

dove la media è presa sulle azioni $a \in \mathcal{A}$ che l'agente può compiere, P_{true} è la distribuzione vera degli esiti e P_{agent} è la distribuzione stimata dall'agente usando il ricordo $\hat{x}$. La distorsione è la divergenza KL media tra le previsioni vere e quelle prodotte usando il ricordo: misura quanto il ricordo $\hat{x}$ degradi la qualità delle previsioni dell'agente rispetto all'evento originale x_e .

C.2 Proprietà della metrica predittiva

La metrica d_{pred} ha tre proprietà rilevanti. Primo, $d_{\text{pred}}(x_e, x_e) = 0$: il ricordo perfettamente fedele non produce distorsione predittiva. Secondo, d_{pred} non è simmetrica in generale: $d_{\text{pred}}(x_e, \hat{x}) \neq d_{\text{pred}}(\hat{x}, x_e)$. Terzo, d_{pred} non dipende dalla distanza percettiva tra x_e e $\hat{x}$ — due ricordi percettivamente diversi dall'originale possono avere la stessa distorsione predittiva se producono previsioni equivalenti, e due ricordi percettivamente simili possono avere distorsioni predittive molto diverse se le differenze percettive ricadono su dimensioni predittivamente critiche.

C.3 La funzione di tasso-distorsione

La funzione di tasso-distorsione nella metrica predittiva è:

$$R_{\text{pred}}(D) = \min_{\{ P(\hat{x} \mid x) : \mathbb{E}[d_{\text{pred}}(x, \hat{x})] \leq D \}} I(X; \hat{X})$$

Per il teorema di Shannon sulla compressione con distorsione, $R_{\text{pred}}(D)$ è decrescente, convessa e continua in D . Il tasso $R_{\text{pred}}(0) = H(X)$ — la sorgente deve essere descritta con la sua entropia completa per distorsione zero — e $R_{\text{pred}}(D) \rightarrow 0$ per $D \rightarrow d_{\text{max}}$, dove d_{max} è la distorsione prodotta dal ricordo costante (ricordare sempre lo stesso evento indipendentemente dall'originale).

C.4 Il lower bound di Shannon-Kolmogorov

La versione algoritmica del teorema di Shannon-Kolmogorov per la compressione lossy afferma che per qualsiasi sorgente x_e e qualsiasi schema di compressione con perdita che produca rappresentazioni di lunghezza $K(m_e)$:

$$\mathbb{E}[d_{\text{pred}}(x_e, \hat{x})] \geq D(K(m_e))$$

dove D è la funzione inversa di R_{pred} . Il lower bound dice che nessuno schema di compressione può garantire distorsione predittiva inferiore a $D(K(m_e))$ se il modello

compresso ha lunghezza KC al più $K(m_e)$. Questo stabilisce il limite fondamentale della memoria: con modelli di KC fissata, esiste una soglia di distorsione predittiva che non può essere abbassata.

Appendice D — Il modello di Nader-LeDoux per il riconsolidamento: formalizzazione

D.1 Il protocollo sperimentale di Nader et al. (2000)

Il riconsolidamento è stato dimostrato sperimentalmente da Nader, Schafe e LeDoux nel 2000 attraverso il seguente protocollo. Ratti condizionati alla paura (fear conditioning) — in cui uno stimolo neutro CS è associato a uno shock elettrico US — vengono esposti di nuovo al CS dopo 24 ore (riattivazione del ricordo). Immediatamente dopo la riattivazione, viene iniettata anisomicina (inibitore della sintesi proteica) nell'amigdala. Se la memoria si trovava in uno stato labile al momento dell'iniezione, l'anisomicina impedisce il riconsolidamento e la memoria viene abolita. Il risultato: i ratti che ricevono anisomicina dopo la riattivazione mostrano amnesia per il condizionamento, mentre i ratti che la ricevono senza riattivazione mantengono la memoria intatta.

D.2 Formalizzazione del processo di labilità e riconsolidamento

Si modella il ciclo di riattivazione-labilità-riconsolidamento come un processo a tre stadi. Nello stadio stabile, m_e è in uno stato di bassa plasticità: gli aggiornamenti di riconsolidamento sono molto piccoli (η efficace ≈ 0). Nello stadio labile, attivato dal richiamo dell'evento, η aumenta bruscamente al valore $\eta_{lab} \gg 0$ per una finestra temporale τ_{lab} . Nel riconsolidamento, m_e viene aggiornato nella direzione dell'errore predittivo δ con tasso η_{lab} , e poi η ritorna al valore basso dello stadio stabile.

La durata della finestra labile τ_{lab} è determinata dalla sintesi proteica che stabilizza le sinapsi aggiornate. Si modella come:

$$\tau_{lab} \propto 1 / (\text{velocità di sintesi proteica})$$

inversamente proporzionale alla velocità con cui le sinapsi vengono re-stabilizzate. Interventi che bloccano la sintesi proteica (anisomicina) aumentano $\tau_{lab} \rightarrow \infty$ ma impediscono la stabilizzazione finale: il modello rimane labile indefinitamente e poi si degrada.

D.3 La finestra di riconsolidamento come vincolo sull'aggiornamento

La finestra τ_{lab} limita la quantità di informazione che può essere incorporata nel riconsolidamento. Se il segnale di errore predittivo δ è complesso — se l'aggiornamento ottimale richiederebbe più tempo di τ_{lab} per essere computato e incorporato — il riconsolidamento è troncato: solo la componente dell'aggiornamento computabile entro τ_{lab} viene incorporata.

Formalmente:

$$\Delta m_e = \eta_{\text{lab}} \cdot \int_0^{\tau_{\text{lab}}} \nabla_{m_e} \delta(t) dt$$

L'integrale è limitato dalla finestra τ_{lab} . Per aggiornamenti con struttura rapida — segnali di errore che si manifestano entro pochi secondi dal richiamo — τ_{lab} è sufficiente. Per aggiornamenti che richiedono elaborazione prolungata — revisioni profonde della struttura del modello — τ_{lab} può essere insufficiente e l'aggiornamento rimane parziale.

Appendice E — Derivazione della curva ROC mnestica e dell'AUC

E.1 Le distribuzioni dei rapporti di verosimiglianza

Si considera la distribuzione di $\log \Lambda(\hat{x}) = \log p(\hat{x} | m_e) - \log p(\hat{x} | p_{\text{sé}})$ sotto le due ipotesi H_0 e H_1 .

Sotto H_0 ($\hat{x}$ generata dalla traccia m_e): la distribuzione di $\log \Lambda$ è spostata verso valori positivi — i ricordi genuini hanno in media $\log p(\hat{x} | m_e) > \log p(\hat{x} | p_{\text{sé}})$. Si assume che $\log \Lambda \mid H_0 \sim N(\mu_0, \sigma_0^2)$ con $\mu_0 > 0$.

Sotto H_1 ($\hat{x}$ generata dalla prior $p_{\text{sé}}$): la distribuzione di $\log \Lambda$ è spostata verso valori negativi. Si assume che $\log \Lambda \mid H_1 \sim N(\mu_1, \sigma_1^2)$ con $\mu_1 < 0$.

E.2 Calcolo dei tassi di errore

Per una soglia τ sulla scala di $\log \Lambda$:

$$\begin{aligned} \alpha(\tau) &= P(\log \Lambda \geq \tau \mid H_1) = 1 - \Phi((\tau - \mu_1) / \sigma_1) \\ \beta(\tau) &= P(\log \Lambda < \tau \mid H_0) = \Phi((\tau - \mu_0) / \sigma_0) \end{aligned}$$

dove Φ è la funzione di distribuzione cumulativa normale standard.

E.3 L'AUC come funzione della separabilità

L'area sotto la curva ROC è:

$$\begin{aligned} \text{AUC} &= P(\log \Lambda(\hat{x}_{\text{genuine}}) > \log \Lambda(\hat{x}_{\text{confab}})) \\ &= \Phi((\mu_0 - \mu_1) / \sqrt{(\sigma_0^2 + \sigma_1^2)}) \\ &= \Phi(d') \end{aligned}$$

dove $d' = (\mu_0 - \mu_1) / \sqrt{(\sigma_0^2 + \sigma_1^2)}$ è la sensibilità discriminativa — la versione gaussiana della misura d' della Signal Detection Theory. AUC cresce con d' , che cresce con la separazione delle medie e decresce con la varianza delle distribuzioni.

La qualità della traccia mnestica m_e determina μ_0 : una traccia ricca sposta la distribuzione dei ricordi genuini verso log Λ alti — μ_0 cresce. La degradazione della traccia abbassa μ_0 verso μ_1 , riducendo d' e l'AUC.

E.4 La soglia ottimale di Bayes

La soglia ottimale nel senso della perdita minima attesa è:

$$\tau^* = (\mu_0 \sigma_1^2 - \mu_1 \sigma_0^2) / (\sigma_0^2 - \sigma_1^2) + \log(C_{\text{FR}} \cdot P(H_1) / C_{\text{FD}} \cdot P(H_0)) \cdot \sigma_0^2 \sigma_1^2 / (\sigma_0^2 - \sigma_1^2)$$

Per il caso semplificato $\sigma_0 = \sigma_1 = \sigma$:

$$\tau^* = (\mu_0 + \mu_1)/2 + \sigma^2 \cdot \log(C_{\text{FR}} \cdot P(H_1) / C_{\text{FD}} \cdot P(H_0)) / (\mu_0 - \mu_1)$$

Il primo termine è il punto di incrocio delle due distribuzioni gaussiane; il secondo termine è lo spostamento dovuto ai costi asimmetrici e alle probabilità a priori. Per $C_{\text{FR}} > C_{\text{FD}}$ (falsi ricordi più costosi dei falsi dimenticamenti), il secondo termine è positivo: la soglia si sposta verso destra, riducendo $\alpha(\tau)$ al costo di aumentare $\beta(\tau)$.

Appendice F — Il modello dell'interpolazione per i falsi ricordi: derivazione

F.1 La distribuzione bimodale e il campionamento

Nel caso di interferenza tra traccia e prior discusso nel Capitolo IV.7, la distribuzione a posteriori sul contenuto dell'evento ha due modi. Per semplificare l'analisi, si considera il caso in cui i due modi sono localizzati intorno a e_{true} (l'evento originale) e e_{prior} (l'evento più a priori probabile per quella cue). La distribuzione bimodale è approssimata da una miscela di gaussiane:

$$p(e \mid c, m_e, p_{\{s\}}) \approx \alpha \cdot \mathcal{N}(e_{\text{true}}, \Sigma_{\text{true}}) + (1-\alpha) \cdot \mathcal{N}(e_{\text{prior}}, \Sigma_{\text{prior}})$$

dove il peso α è:

$$\alpha = \sigma(\log p(m_e | e_{\text{true}}) - \log p(e_{\text{prior}} | p_{\text{sc}}))$$

la funzione sigmoide applicata alla differenza tra il log-likelihood della traccia e il log-likelihood della prior. $\alpha \in (0,1)$ e cresce con la forza della traccia e decresce con la dominanza della prior.

F.2 Il ricordo come campione dalla miscela

Il ricordo generato $\hat{x}$ è ottenuto campionando dalla distribuzione bimodale. Condizionatamente al campionamento da uno dei due modi, si ha:

$$\begin{aligned}\hat{x} | (\text{campione dal primo modo}) &\sim \mathcal{N}(x_{\{e_{\text{true}}\}}, \Sigma_{\text{true}}) \\ \hat{x} | (\text{campione dal secondo modo}) &\sim \mathcal{N}(x_{\{e_{\text{prior}}\}}, \Sigma_{\text{prior}})\end{aligned}$$

Marginalizzando sull'assegnazione al modo:

$$\begin{aligned}\mathbb{E}[\hat{x}] &= \alpha \cdot x_{\{e_{\text{true}}\}} + (1-\alpha) \cdot x_{\{e_{\text{prior}}\}} \\ \text{Var}[\hat{x}] &= \alpha(1-\alpha)(x_{\{e_{\text{true}}\}} - x_{\{e_{\text{prior}}\}})^2 + \alpha \cdot \Sigma_{\text{true}} + (1-\alpha) \cdot \Sigma_{\text{prior}}\end{aligned}$$

Il valore atteso del ricordo è l'interpolazione lineare tra i due eventi — l'evento genuino e l'evento confabulato — con pesi α e $(1-\alpha)$. La varianza ha un termine aggiuntivo $\alpha(1-\alpha)(x_{\{e_{\text{true}}\}} - x_{\{e_{\text{prior}}\}})^2$ che misura la variabilità dovuta all'incertezza su quale modo è stato campionato: è massima quando $\alpha = 0.5$ (massima incertezza tra i due modi) e si annulla quando $\alpha \rightarrow 0$ o $\alpha \rightarrow 1$ (un modo domina completamente).

F.3 Previsioni sulla struttura dei falsi ricordi

Il modello dell'interpolazione produce previsioni specifiche. I dettagli del falso ricordo che deviano dall'evento originale devono deviare nella direzione di e_{prior} — nella direzione dell'evento più a priori plausibile per quella cue. La deviazione non è arbitraria: ha la forma specifica di $x_{\{e_{\text{prior}}\}} - x_{\{e_{\text{true}}\}}$, pesata da $(1-\alpha)/\alpha$.

Se α è vicino a 1 (traccia forte), le deviazioni sono piccole — il ricordo è vicino all'originale con piccole distorsioni verso e_{prior} . Se α è vicino a 0 (prior dominante), il ricordo è quasi completamente determinato da e_{prior} con piccole tracce dell'evento originale. Nei casi intermedi, il ricordo è un'interpolazione: coerente con entrambi gli eventi in certi dettagli, incoerente con ciascuno in altri.

Appendice G — Convergenza dell'experience replay e riduzione dell'interferenza retroattiva

G.1 Il problema dell'interferenza retroattiva come ostacolo all'ottimizzazione

Si formalizza l'interferenza retroattiva come la degradazione della funzione di perdita su eventi precedenti causata dall'aggiornamento su un nuovo evento. Sia $\mathcal{L}_e$ il valore della funzione obiettivo per l'evento e prima dell'aggiornamento su un nuovo evento e' . Dopo l'aggiornamento $\nabla_{m_{e'}} J(m_{e'})$:

$$\mathcal{L}_{e'}^{\text{post}} = \mathcal{L}_{e'}^{\text{pre}} + \langle \nabla_m \mathcal{L}_e, \eta \cdot \nabla_m J(m_{e'}) \rangle + O(\eta^2)$$

L'interferenza retroattiva è il secondo termine — il prodotto interno tra il gradiente della perdita su e e il gradiente dell'aggiornamento su e' . Se i due gradienti sono allineati, l'aggiornamento migliora entrambi gli eventi. Se sono ortogonali, non c'è interferenza. Se sono anti-allineati, l'aggiornamento su e' peggiora la rappresentazione di e .

G.2 L'experience replay come ortogonalizzazione dei gradienti

L'experience replay riduce l'interferenza retroattiva approssimando il gradiente della funzione obiettivo globale anziché quello individuale. Un aggiornamento su un minibatch $B = \{e_1, \dots, e_b\}$ ha la forma:

$$\nabla_m J_{\text{global}}(B) = (1/b) \cdot \sum_{\{e_i \in B\}} \nabla_m J(m_{\{e_i\}})$$

La media dei gradienti individuali tende a cancellare le componenti anti-allineate — le componenti idiosincratiche degli aggiornamenti individuali che producono interferenza — e a preservare le componenti allineate — quelle che riflettono la struttura comune agli eventi. Quando $b \rightarrow N$ (minibatch di dimensione pari all'intero set di eventi), il gradiente converge al gradiente della perdita media su tutti gli eventi, che è il gradiente che minimizza J_{global} senza interferenza.

G.3 Il guadagno di convergenza quantificato

Per un problema di ottimizzazione L -smooth con varianza del gradiente σ^2_{grad} , il numero di iterazioni necessarie per raggiungere una ϵ -vicinanza del minimo è:

$$T_{\text{online}} = O(L \cdot \sigma^2_{\text{grad}} / \epsilon) \quad (\text{aggiornamento online, } b=1)$$

$$T_{\text{replay}} = O(L \cdot \sigma^2_{\text{grad}} / (b \cdot \epsilon)) \quad (\text{experience replay, minibatch } b)$$

Il guadagno è un fattore b nella velocità di convergenza. Ma il vantaggio del replay non è solo la velocità: è l'accessibilità del minimo globale. L'aggiornamento online converge al minimo della funzione di perdita individuale $J(m_e)$, che non coincide con il minimo di J_{global} quando i gradienti degli eventi sono correlati. Il replay converge al minimo di J_{global} indipendentemente dalla correlazione dei gradienti, perché la media su B rende il gradiente effettivo ortogonale alle componenti di interferenza.

Bibliografia ragionata

La bibliografia è organizzata in nuclei tematici corrispondenti agli assi principali della monografia. Per ciascun testo è indicato il ruolo specifico nell'argomentazione e, dove necessario, la distanza critica che la monografia mantiene rispetto alle sue tesi.

I. Fondamenti neurobiologici: riconsolidamento e consolidamento sistemico

Nader, K., Schafe, G. E., & LeDoux, J. E. (2000). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature*, 406(6797), 722–726.

Il testo che ha rilanciato lo studio del riconsolidamento nella forma sperimentalmente rigorosa che il campo moderno riconosce. Il protocollo ippocampo-amigdala-anisomicina è il punto di partenza della formalizzazione del Capitolo II e dell'Appendice D. Da notare: Nader e LeDoux hanno successivamente raffinato la propria posizione sulla generalità del riconsolidamento — non tutti i ricordi si riconsolidano in ogni circostanza — e la monografia incorpora questa raffinazione nella modulazione del tasso η per valenza emotiva.

Misanin, J. R., Miller, R. R., & Lewis, D. J. (1968). Retrograde amnesia produced by electroconvulsive shock after reactivation of a consolidated memory trace. *Science*, 160(3827), 554–555.

La scoperta originale del riconsolidamento, rimasta ignorata per trent'anni. Misanin, Miller e Lewis mostrano che un ricordo consolidato può essere reso nuovamente vulnerabile all'amnesia se riattivato prima del trattamento amnesico. Il meccanismo è lo stesso di Nader et al. ma su un substrato sperimentale diverso. Rilevante come evidenza della generalità del fenomeno attraverso paradigmi sperimentali distinti.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3), 419–457.

Il framework dei Complementary Learning Systems (CLS), fondamentale per il Capitolo VII. McClelland, McNaughton e O'Reilly dimostrano formalmente che un singolo sistema di apprendimento non può risolvere simultaneamente il problema dell'acquisizione rapida di nuovi episodi e quello della generalizzazione lenta di strutture condivise: i due obiettivi richiedono tassi di apprendimento incompatibili. Il sistema ippocampo-corteccia è la soluzione biologica a questo trade-off computazionale. L'Appendice G formalizza la riduzione dell'interferenza retroattiva che il framework predice.

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences*, 20(7), 512–534.

L'aggiornamento del framework CLS alla luce dei progressi nelle neuroscienze e nell'apprendimento automatico. I paralleli tra experience replay nei sistemi di reinforcement learning e il consolidamento sistemico durante il sonno sono discussi esplicitamente. L'algoritmo di DQN di DeepMind — che usa experience replay per stabilizzare l'apprendimento — è presentato come un'istanza artificiale del meccanismo biologico. La connessione con il Capitolo VII è diretta.

Stickgold, R. (2005). Sleep-dependent memory consolidation. *Nature*, 437(7063), 1272–1278.

La rassegna più citata sul consolidamento durante il sonno. Stickgold sintetizza le evidenze che distinguono le funzioni del sonno REM (generalizzazione, integrazione emotiva, simulazione generativa) e del sonno SWS (consolidamento sistemico, trasferimento ippocampo-corteccia). Essenziale per la distinzione delle due fasi proposta nel Capitolo VII.6.

Walker, M. P., & Stickgold, R. (2006). Sleep, memory, and plasticity. *Annual Review of Psychology*, 57, 139–166.

Il testo più completo sulla relazione tra sonno e plasticità sinaptica. Documenta il sleep-dependent generalization — il miglioramento della capacità di estrarre strutture statistiche dopo il sonno — che il Capitolo VII.5 interpreta come minimizzazione della coerenza inter-evento durante l'experience replay.

II. Ricostruzione mnestica, falsi ricordi e schema

Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press.

Il testo fondativo della tradizione ricostruttiva. Bartlett mostra che il ricordo di storie culturalmente distanti viene sistematicamente ricostruito nella direzione degli schemi familiari al soggetto. Il concetto di schema — la struttura cognitiva che orienta la ricostruzione — è l'antenato diretto del programma del sé $p_{\{sé\}}$ della monografia. Il Capitolo III si costruisce in dialogo diretto con Bartlett, sostituendo la nozione qualitativa di schema con la formalizzazione del codec narrativo.

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology*, 7(4), 560–572.

Il primo studio sistematico sull'effetto della domanda suggestiva sui ricordi testimoniati. Loftus dimostra che basta modificare un verbo in una domanda — "si sono scontrate" vs "si sono toccate" — per modificare le stime quantitative dei testimoni e impiantare falsi ricordi di dettagli non presenti nell'evento originale. Il modello dell'interpolazione dell'Appendice F è la formalizzazione del meccanismo che Loftus documenta empiricamente.

Loftus, E. F., & Pickrell, J. E. (1995). The formation of false memories. *Psychiatric Annals*, 25(12), 720–725.

Lo studio del "lost in the mall" — il primo esperimento che dimostra l'impiantazione di ricordi completamente falsi di eventi mai accaduti, attraverso suggestione narrativa. Il 25% dei soggetti sviluppa ricordi vividi di un'esperienza di smarrimento infantile in un centro commerciale mai avvenuta. Questo risultato è la motivazione empirica centrale per il modello bayesiano del Capitolo IV: la confabulazione non produce solo distorsioni di dettagli ma istanze complete dalla prior.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding*. Lawrence Erlbaum.

L'introduzione della nozione di script come struttura cognitiva che codifica le sequenze tipiche di eventi in contesti routinari. Gli script sono il correlato cognitivo degli schemi narrativi del Capitolo III: la distorsione di completezza causale che elimina gli eventi che non si inseriscono nello script è documentata sistematicamente in studi ispirati da questo framework. La monografia usa la nozione di script come caso paradigmatico di prior forte (basso ConfScore) che produce alta confabulazione.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(4), 803–814.

Il paradigma DRM (Deese-Roediger-McDermott): i soggetti presentati con liste di parole associate a un tema (sonno, riposo, stanchezza, cuscino...) ricordano frequentemente la parola critica (sonno) anche se non era nella lista. Il fenomeno è il correlato sperimentale diretto del campionamento dalla prior: la parola critica ha ConfScore altissimo — è l'evento più a priori plausibile dato l'insieme delle cue — ed è sistematicamente confabulata. Il Capitolo IV.6 usa questo paradigma come prototipo per la misura di ConfScore.

III. Identità personale e continuità psicologica

Locke, J. (1689/1975). *An Essay Concerning Human Understanding* (P. H. Nidditch, Ed.). Oxford University Press. (Libro II, Capitolo XXVII.)

Il locus classicus della teoria della continuità mnestica come fondamento dell'identità personale. Locke argomenta che la persona è definita dalla continuità della coscienza e della memoria: essere la stessa persona attraverso il tempo significa ricordare le proprie esperienze passate. La monografia parte da Locke per mostrare che, se la continuità mnestica è soggetta alla deriva inevitabile, la continuità dell'identità personale è strutturalmente una questione di grado piuttosto che di tutto-o-niente.

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press. (Parte III.)

Parfit argomenta contro la tesi lockiana che l'identità personale sia quello che conta: la sopravvivenza psicologica — la continuità delle credenze, dei desideri, delle intenzioni — è ciò che importa, indipendentemente dall'identità numerica. Il concetto di continuità psicologica di Parfit è formalmente più debole della continuità mnestica di Locke, e la misura $\text{Cont}(t_0, T)$ del Capitolo X.4 è una formalizzazione della continuità psicologica nel senso di Parfit, non dell'identità numerica nel senso di Locke.

Schechtman, M. (1996). *The Constitution of Selves*. Cornell University Press.

Schechtman propone la *narrative self-constitution view*: l'identità personale non è costituita dalla continuità di particolari stati mentali, ma dalla capacità di incorporare le proprie esperienze in una narrazione autobiografica coerente. La monografia è in dialogo diretto con questa posizione: il codec narrativo del Capitolo III è la formalizzazione del meccanismo che Schechtman descrive, e i limiti del codec (le distorsioni sistematiche) sono i limiti della narrative self-constitution view come teoria dell'identità.

IV. Memoria collettiva e politica della memoria

Halbwachs, M. (1950). *La mémoire collective*. Presses Universitaires de France.

Il testo fondativo dello studio sociologico della memoria collettiva. Halbwachs argomenta che tutta la memoria è socialmente mediata: anche i ricordi individuali sono costruiti attraverso quadri sociali di riferimento — i valori, le norme, le narrazioni del gruppo di appartenenza. La monografia formalizza questa tesi nel modello della rete mnestica collettiva del Capitolo V, dove la prior individuale $p(e \mid p_{\text{sé}}^{(i)})$ è parzialmente determinata dalla prior collettiva attraverso il riconsolidamento mediato dagli scambi comunicativi.

Assmann, J. (1992). *Das kulturelle Gedächtnis: Schrift, Erinnerung und politische Identität in frühen Hochkulturen*. C. H. Beck.

Assmann distingue la *memoria comunicativa* — il ricordo degli eventi degli ultimi tre o quattro generazioni, trasmesso informalmente — dalla *memoria culturale* — il ricordo fondativo della tradizione, stabilizzato attraverso riti, testi canonici e specialisti della memoria. La distinzione corrisponde, nel modello della monografia, alla differenza tra aggiornamento bilaterale tra pari (memoria comunicativa) e aggiornamento istituzionale asimmetrico (memoria culturale). La monografia formalizza questa distinzione nei termini della topologia della rete G e dei tassi $\eta_{-}\{ij\}$ vs $\eta_{-}\{0i\}$.

Olick, J. K. (1999). Collective memory: The two cultures. *Sociological Theory*, 17(3), 333–348.

Olick distingue due approcci alla memoria collettiva: quello che la intende come somma dei ricordi individuali e quello che la intende come insieme di pratiche culturali pubbliche. La formalizzazione del Capitolo V prende una posizione precisa: la memoria collettiva è la distribuzione empirica $M_e^{\{coll\}}$ dei modelli individuali dotata della topologia di rete — non è né la loro somma né è riducibile alle pratiche istituzionali, ma emerge dall'interazione tra i due livelli.

Young, J. E. (1993). *The Texture of Memory: Holocaust Memorials and Meaning*. Yale University Press.

Lo studio più ricco sull'architettura della memoria collettiva attraverso i memoriali dell'Olocausto. Young analizza come le scelte architettoniche — la forma del monumento, il suo posizionamento, le iscrizioni — non siano neutre ma orientino la deriva della memoria collettiva nella direzione di specifiche interpretazioni dell'evento. Nel quadro del modello, ogni monumento è un'istanziatura fisica del modello canonico $m_e^{\{0\}}$ dell'istituzione mnestica, con $\eta_{-}\{0i\}$ tanto più alto quanto più il memoriale è culturalmente centrale e fisicamente prominente.

V. Confabulazione clinica e testimonianza oculare

Kopelman, M. D. (1987). Two types of confabulation. *Journal of Neurology, Neurosurgery, and Psychiatry*, 50(11), 1482–1487.

Kopelman distingue la confabulazione *spontanea* — produzione continua di falsi ricordi senza stimolazione esterna — dalla confabulazione *provocata* — falsi ricordi prodotti in risposta a domande. La distinzione è rilevante per il modello bayesiano: la confabulazione spontanea corrisponde al campionamento dalla prior in assenza di cue esterna (la distribuzione a posteriori è dominata dalla prior in ogni circostanza); la confabulazione provocata corrisponde al campionamento dominato dalla prior in risposta a una cue (la cue attiva la prior anziché la traccia). La monografia usa la terminologia di Kopelman implicitamente nella distinzione tra ConfScore alto in

assenza di cue (confabulazione spontanea) e alto in presenza di cue (confabulazione provocata).

Berlyne, N. (1972). Confabulation. *British Journal of Psychiatry*, 120(554), 31–39.

Un'analisi sistematica dei contenuti delle confabulazioni in pazienti con sindrome di Korsakoff. Berlyne documenta che i contenuti confabulati sono quasi invariabilmente eventi plausibili per il paziente — routine, persone familiari, luoghi consueti — e raramente eventi bizzarri o fantastici. Questa osservazione è il correlato clinico della previsione del Capitolo IV.5: la confabulazione campiona dalla prior, e la prior è la distribuzione degli eventi tipici per quell'agente.

Wells, G. L., & Olson, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology*, 54(1), 277–295.

La rassegna più completa sulla testimonianza oculare come problema cognitivo e legale. Wells e Olson documentano sistematicamente i fattori che degradano l'AUC della curva ROC mnestica del testimone — stress, distanza, brevità del contatto, condizioni di illuminazione — e quelli che modificano la soglia τ — pressione dell'esaminatore, formulazione delle domande, struttura della lineup. La formalizzazione del Capitolo IX traduce le loro conclusioni empiriche in previsioni quantitative sul trade-off α - β .

Innocence Project. (2020). *Eyewitness Misidentification*. Disponibile su: <https://innocenceproject.org/eyewitness-misidentification/>

Il corpus documentale che mostra che circa il 70% dei casi esonerati per evidenza del DNA contenevano testimonianze oculari errate. Il Capitolo IX.6 usa questi dati come evidenza che la soglia evolutivamente calibrata τ^* del sistema mnestico umano produce tassi di falsi ricordi sistematicamente troppo alti per il contesto giudiziario. Non un testo accademico ma una fonte empirica primaria per la frequenza reale degli errori di testimonianza.

VI. Memoria protesica e archivi digitali

Henkel, L. A. (2014). Point-and-shoot memories: The influence of taking photos on memory for a museum tour. *Psychological Science*, 25(2), 396–402.

Lo studio originale del photo-taking impairment effect. Henkel mostra che fotografare oggetti in un museo ne riduce il successivo ricordo rispetto al semplice osservarli, e che l'effetto è parzialmente mitigato quando i soggetti fotografano con zoom su un dettaglio specifico (che richiede elaborazione più profonda). La derivazione formale nell'Appendice C della monografia è costruita direttamente a partire da questo risultato empirico.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333(6043), 776–778.

Lo studio che documenta l'"effetto Google": sapere che un'informazione sarà recuperabile in futuro riduce la probabilità di memorizzarla internamente. I soggetti informati che le risposte alle domande saranno salvate nel computer mostrano peggiore memoria per le risposte ma migliore memoria per dove troveranno le risposte. Questo è un caso di ottimizzazione della codifica sotto il vincolo della protetica disponibile: il sistema riduce $KC(m_e)$ perché sa che e_e è disponibile, esattamente come predice il modello dell'Appendice C.

Landsberg, A. (2004). *Prosthetic Memory: The Transformation of American Remembrance in the Age of Mass Culture*. Columbia University Press.

Il testo che introduce il termine "memoria protetica" nel contesto degli studi culturali. Landsberg analizza come i media di massa — cinema, televisione, musei — producano esperienze mnestiche vivide di eventi storici non vissuti personalmente. La monografia usa il termine in senso più ampio, includendo le tracce digitali personali, e la formalizza nella funzione di tasso-distorsione con protetica dell'Appendice C.

Wegner, D. M., Giuliano, T., & Hertel, P. (1985). Cognitive interdependence in close relationships. In W. Ickes (Ed.), *Compatible and Incompatible Relationships* (pp. 253–276). Springer.

Il concetto di *transactive memory* — la memoria distribuita tra i membri di un gruppo, in cui ciascuno specializza la propria memoria in un sottoinsieme di domini e si affida agli altri per i domini complementari. La transactive memory è il precursore teorico della formalizzazione della memoria collettiva come sistema distribuito del Capitolo V, e anticipa il modello della rete mnestica con pesi η_{ij} differenziati per relazione.

VII. Signal Detection Theory e metacognizione mnestica

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. Wiley.

Il testo fondativo della Signal Detection Theory applicata alla psicologia. La misura d' e la costruzione della curva ROC sono sviluppate qui nel contesto della percezione e poi estese alla memoria da Murdock (1974). L'Appendice E della monografia segue la derivazione di Green e Swets per la curva ROC mnestica, adattando la struttura della distribuzione gaussiana al caso del rapporto di verosimiglianza mnestico $\Lambda(\hat{x})$.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, 26, 125–173.

Il framework metacognitivo per la memoria. Nelson e Narens distinguono il livello-oggetto (i processi mnestici) dal livello-meta (il monitoraggio e il controllo dei processi mnestici). La correlazione Gamma — la misura di correlazione tra la fiducia

soggettiva e l'accuratezza oggettiva — è introdotta qui come misura di calibrazione metacognitiva. Il Capitolo IX.7 interpreta la correlazione Gamma come misura di quanto la soglia individuale τ sia calibrata rispetto all'AUC individuale della curva ROC.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126(4), 349–370.

Koriat analizza come i soggetti usino cue di fluenza, familiarità e coerenza per monitorare la qualità dei propri ricordi. Le cue di fluenza — la facilità con cui un'informazione viene generata — sono correlate positivamente con la fiducia soggettiva ma non necessariamente con l'accuratezza oggettiva. Questo è il meccanismo sottostante al paradosso del ricordo più vivo del vissuto del Capitolo X.5: la fluenza generativa aumenta la vivacità soggettiva $V(\hat{x})$ indipendentemente dalla fedeltà alla traccia originale.

VIII. Principio di energia libera e predictive coding

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

La formulazione più accessibile del principio di energia libera. Friston argomenta che i sistemi biologici minimizzano la sorpresa — tecnicamente l'energia libera variazionale — come principio organizzativo unificato della cognizione, della percezione e dell'azione. L'Appendice A della monografia segue la derivazione di Friston e stabilisce la connessione formale con la funzione obiettivo mnestica $J(m_e)$.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

Il testo più accessibile sul predictive coding e sul framework della mente predittiva. Clark sintetizza le evidenze neuroscientifiche e computazionali per l'idea che il cervello sia essenzialmente un sistema predittivo che minimizza l'errore di previsione a tutti i livelli della gerarchia cognitiva. La monografia usa il framework di Clark come contesto per la funzione obiettivo del sistema mnestico, senza però adottare la sua posizione più forte — che il predictive coding sia l'unico principio organizzativo del cervello.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), 79–87.

Il modello computazionale che ha introdotto il predictive coding nella neuroscienze computazionali. Rao e Ballard dimostrano che un modello gerarchico in cui ogni livello predice l'attività del livello inferiore e propaga solo il residuo di errore replica

molte proprietà elettrofisiologiche della corteccia visiva. Il modello è il precursore computazionale del principio di energia libera di Friston, e la sua struttura gerarchica è il prototipo del sistema generativo che la monografia descrive a livello mnestico.